Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Modelos Estatísticos para a Previsão de Inactividade de Pré-Pagos

Marta Gonçalves Cruces Simão Portugal

Dissertação
Versão Pública

Mestrado em Estatística e Investigação Operacional
Especialização em Estatística

2012- 2013

Universidade de Lisboa
Faculdade de Ciências
Departamento de Estatística e Investigação Operacional

Modelos Estatísticos para a Previsão de Inactividade de Pré-Pagos

Marta Gonçalves Cruces Simão Portugal

Dissertação
Mestrado em Estatística e Investigação Operacional
Especialização em Estatística

Orientadores:
Professora Dra. Marília Antunes
Dra. Paula Figueiredo Mestres

2012 - 2013

Agradecimentos

A realização desta dissertação marca o fim de uma importante etapa da minha vida.

É com uma grande satisfação que expresso aqui o meu mais profundo agradecimento a todos aqueles que tornaram possível a realização deste projecto. No entanto, há contributos de natureza diversa que não podem e nem devem deixar de ser realçados. Assim sendo, gostaria de agradecer a todos aqueles cujo contributo foi essencial para a sua conclusão:

À minha orientadora, Prof. Dra. Marília Antunes, um muito obrigada pelo constante apoio, incentivo e ensinamentos que tornaram este trabalho numa realidade, assim como pelas críticas, correcções e sugestões sempre cheias de carinho.

À minha co-orientadora, Dra. Paula Figueiredo, pela oportunidade, ajuda e orientação fornecida ao longos destes meses e por ter acreditado sempre em mim.

A todos os elementos da empresa em questão, pela disponibilidade, pela ajuda e por me terem feito sentir em casa.

Aos meus colegas de mestrado, pelos momentos de entusiasmo partilhados em conjunto.

À Mariana e ao Sandro por tudo o que passamos dentro (e fora) destas quatro paredes e, acima de tudo, por serem amigos verdadeiros.

À minha familia, em particular ao meu namorado Adrià, à minha mãe, ao Rui e aos meus irmãos pelo suporte emocional e ajuda incansável. Sem vocês não teria sido possível.

Marta

Conteúdo

1 Introdução **3**
- 1.1 Objectivos e Metodologia 4
- 1.2 Estrutura do trabalho 5

2 Análise exploratória dos dados **8**
- 2.1 A população, a amostra e a base de dados 8
- 2.2 As variáveis em estudo 9
- 2.3 Abordagens ao problema 13

3 O modelo linear generalizado **29**
- 3.1 O modelo de regressão logística 32
  - 3.1.1 Ajustamento do modelo 33
  - 3.1.2 Coeficientes do modelo 35
  - 3.1.3 Diagnóstico do modelo 37

4 Aplicação **44**
- 4.1 Modelação da amostra completa 44
  - 4.1.1 Estratégias de modelação 44
  - 4.1.2 Diagnóstico do modelo 45
- 4.2 Modelação da amostra censurada 60
  - 4.2.1 Estratégias de modelação 61
  - 4.2.2 Diagnóstico do modelo 61
- 4.3 Simulação da utilização do modelo 73

5 Conclusões e trabalho futuro **80**

Lista de Figuras

2.1 Diferenças nos valores da variável A para as diferentes categorias das variáveis B, C, F e G. 11
2.2 Diferenças nos valores das variáveis M e N para as diferentes categorias das variáveis B, C, F e G. 12
2.3 Diferenças nos valores da variável A, em m_1, entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012. 14
2.4 Diferenças da média dos valores da variável M entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012. 15
2.5 Diferenças da média dos valores da variável N entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012. 16
2.6 Diferenças nos valores da variável B entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012. 17
2.7 Diferenças da média dos valores da variável *mean* M entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012. 19
2.8 Diferenças da média dos valores da variável *mean* N entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012. 21
2.9 Categorias de comportamento para um serviço j, no caso em que $x_j - s_j > 0$. 23
2.10 Categorias de comportamento para um serviço j, no caso em que $x_j - s_j \leq 0$. 24

3.1 Exemplo de uma curva ROC. 43

4.1 Resíduos padronizados do modelo II. 50
4.2 Resíduos *versus* valores ajustados do modelo II. 51

4.3 Normal Q-Q plot para o modelo II. 52
4.4 Histograma dos resíduos do modelo II. 53
4.5 Scale-location plot do modelo II. 54
4.6 Resíduos vs *leverage*. 55
4.7 Distância de Cook para o modelo II. 56
4.8 Distância de Cook vs *leverage* para o modelo II. 57
4.9 Curva ROC do modelo II. 59
4.10 Acertos, sensibilidade, especificidade e performance do modelo II. . . . 60
4.11 Resíduos padronizados do modelo IV. 65
4.12 Resíduos *versus* valores ajustados do modelo IV. 66
4.13 Normal Q-Q plot para o modelo IV. 67
4.14 Histograma dos resíduos do modelo IV. 68
4.15 Scale-location plot do modelo IV. 69
4.16 Resíduos *versus leverage* para o modelo IV. 70
4.17 Distância de Cook para o modelo IV. 71
4.18 Distância de Cook vs *leverage* para o modelo IV. 72
4.19 Curva ROC do modelo IV. 74
4.20 Acertos, sensibilidade, especificidade e performance do modelo IV. . . . 75

Lista de Tabelas

2.1 Principais medidas de localização das variáveis retiradas da base de dados, em Outubro de 2012. 10
2.2 Principais medidas de localização da variável *mean M*, para serviços inactivos I em Outubro de 2012 e para serviços activos I na mesma data, respectivamente. 14
2.3 Principais medidas de localização da variável *mean N*, para serviços inactivos I em Outubro de 2012 e para serviços activos I na mesma data, respectivamente. 15
2.4 Principais medidas de localização da variável M, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente. 18
2.5 Principais medidas de localização da variável N, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente. 18
2.6 Principais medidas de localização da variável *mean M*, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente. 20
2.7 Principais medidas de localização da variável *mean N*, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente. 20
2.8 Correlação entre as variáveis A, B e C. 25
2.9 Principais medidas de localização os limites de controlo inferior para as variáveis A, B e C, respectivamente. 27

3.1 Matriz de confusão para um modelo de regressão logística. 41

4.1 *Summary* do modelo I. 46
4.2 *Step* do modelo I. 47
4.3 *Summary* do modelo II. 48

4.4 *Variance Inflation Factors* das covariáveis do modelo II. 49
4.5 *Summary* do modelo III. 62
4.6 *Step* do modelo III. 63
4.7 *Summary* do modelo IV. 64
4.8 *Variance Inflation Factors* das covariáveis do modelo IV. 65
4.9 Tabela de predições categorizadas do modelo II. 76
4.10 Matriz de probabilidades de transição a um mês. 78
4.11 Matriz de probabilidades de transição a dois meses. 78
4.12 Matriz de probabilidades de transição a três meses. 79

Resumo

O aumento da competitividade global tem acarretado um inevitável aumento de custos associados à angariação de novos consumidores, impondo às empresas o desafio de mudar a sua forma de actuar perante os seus, cada vez mais exigentes, consumidores a fim de evitar a sua perda. As empresas de telecomunicações não são excepção, uma vez que operam num mercado cada vez mais saturado e exigente, no qual as mudanças de operador são frequentes e facilitadas. Perante tal desafio, este trabalho visa aprofundar o estudo do comportamento individual do cliente, identificando os factores decisivos para o decréscimo do comportamento, assim como propor uma estratégia de predição e prevenção desse decréscimo. Neste estudo, a técnica utilizada para a classificação dos clientes foi a regressão logística, combinada com a metodologia de cadeias de Markov. Esta dissertação de mestrado está inserida num projecto profissional na área das telecomunicações.

Keywords: mercado das telecomunicações, *churn*, retenção de clientes, modelos lineares generalizados, regressão logística, cadeias de Markov.

Abstract
The increase of the global competition has caused an inevitable rise in the costs associated to the acquisition of new customers, imposing companies the challenge of changing their way to act regarding their, increasingly more demanding, costumers, so that they are able to avoid their loss. Telecommunication companies are no exception to this phenomenon, since they are operating in a progressively more saturated and demanding market, in which operator changes are frequent and easy. Given that challenge, this dissertation aims to deepen the study of the individual behavior of the client (regarding usage), identify the crucial factors that lead to usage decrease, as well as propose a strategy for prediction and prevention of this decrease. In this study, the classification technique used was the logistic regression combined with the Markov chains methodology. This project regards the telecommunication field.

Keywords: telecommunication industry, churn,client retention, generalized linear models, logistic regression, Markov chains.

Capítulo 1

Introdução

No final do século XX e início do século XXI, as telecomunicações transformaram-se num dos sectores mais dinâmicos e emergentes da economia mundial, sendo que se estima que em 2010 aproximadamente cinco biliões de pessoas no Mundo possuía um telemóvel.

Os aparelhos e o leque de serviços disponibilizados pelos mesmos (desde transmissão de voz e dados a câmaras digitais e e-mail) tornaram-se essenciais para a grande maioria da população, podendo-se dizer que as telecomunicações móveis já alcançaram o estatuto de bem essencial do século XXI.

Com o crescimento do número de clientes nos últimos anos, com a globalização do mercado das telecomunicações e com o aumento do número de operadores concorrentes, não é de estranhar que a indústria das telecomunicações se foque, cada vez mais, em mais do que apenas fornecer um serviço de comunicação móvel. Este crescimento criou nos operadores a necessidade e a preocupação de apostar na qualidade do serviço, uma vez que esta é a principal forma de garantir quer a lealdade por parte do consumidor, quer a competitividade na aquisição de novos clientes, quer a retenção dos clientes já existentes. No entanto, os rápidos e constantes avanços tecnológicos suscitam no consumidor a vontade de mudança, quer na procura de melhores condições financeiras, quer na procura de equipamentos mais sofisticados.

Num mercado saturado como aquele em que nos encontramos, e uma vez que o custo associado à aquisição de novos clientes é muito superior ao custo associado à sua retenção (Richter (2010)) a predição de *churn* (acto de um cliente abandonar o operador actual em favor de um operador concorrente)

tornou-se cada vez mais imperativa, sendo visível o crescente esforço das operadoras neste sentido.

Existem dois tipos de *churn*:

- O *churn* voluntário que ocorre quando um consumidor, por vontade própria, termina os serviços com a operadora, quer por razões deliberadas (razões relacionadas com o serviço da operadora), quer por razões acidentais (razões que fogem ao controlo do consumidor, como mudança de residência ou problemas financeiros pessoais).
- O *churn* involuntário que é resultado de uma acção da própria operadora, na qual se dá a cessação do contrato com o consumidor (por falta de pagamentos, por exemplo).

Todos estes conceitos fazem sentido (e são aprofundadamente estudados) num contexto de clientes com serviços pós-pagos, uma vez que a cessação de contrato (ou serviço) tem de ser formalizada através de um contacto, quer do consumidor com a operadora (*churn* voluntário), quer da operadora com o consumidor (*churn* involuntário).

No entanto, quando se trata de clientes com serviços pré-pagos (doravante designados por clientes pré-pagos), pode-se questionar a importância do conceito de *churn* uma vez que o contacto entre o consumidor e a operadora na maioria das vezes não existe (o consumidor simplesmente deixa de utilizar o serviço fornecido pela operadora) tornando mais difícil (por vezes mesmo impossível) o contacto da operadora com o consumidor, quando o *churn* é detectado.

Neste contexto, introduz-se a noção de comportamento (e consequentemente de quebra do mesmo) como o conceito chave quando o objectivo é a retenção de clientes pré-pagos.

1.1 Objectivos e Metodologia

Esta dissertação visa aprofundar o estudo do fenómeno *churn* na classe de consumidores de voz móvel pré-pagos, sob a forma de estudo do seu comportamento, de forma a que a sua detecção seja feita na fase mais inicial

possível, permitindo uma abordagem proactiva na fase da retenção de clientes.

Para tal foi utilizada informação mensal agregada relativa aos últimos seis meses de clientes pré-pagos de voz móvel activos à data de Novembro de 2012. Para estes clientes foram analisadas as diversas variáveis relativas ao comportamento e foram seleccionadas as consideradas relevantes e representativas, usando os métodos estatísticos considerados relevantes. Foi elaborada uma definição formal de comportamento, assim como de quebra do mesmo e foi ajustado um modelo de regressão logística para prever se um consumidor iria quebrar o seu comportamento ou não, no mês seguinte ao mês corrente. Foi simulada a utilização do modelo (com dados reais) durante sete meses consecutivos tendo como objectivo criar uma sucessão de valores de propensão (definindo-se propensão como o valor da variável resposta do modelo de regressão logística ajustado), as quais foram categorizadas e a cujas categorias foi aplicada a metodologia de cadeias de Markov de forma a calcular matrizes de probabilidade de transição a um passo, dois passos e três passos.

Como alternativa foi ajustado um outro modelo de regressão logística, com a mesma variável resposta, tendo como base a amostra sem ter em conta os clientes que não apresentavam comportamento significativo há pelo menos cinco meses. O objectivo da construção deste último modelo é retirar a influência dos clientes que apresentavam comportamento residual na estimação dos parâmetros do primeiro modelo.

É de frisar que todos os dados utilizados foram agregados e anonimizados, não havendo forma de relacionar os dados em estudo com o cliente ao qual pertencem. Todos os métodos estatísticos foram realizados utilizando o software estatístico R.

1.2 Estrutura do trabalho

Como em todos os projectos, nem todas as ideias apresentadas foram continuadas (algumas por inadequabilidade aos dados, outras por inadequabilidade aos objectivos da empresa), no entanto há um fio condutor que as liga, fazendo com que a sua apresentação e discussão torne mais fácil seguir a sequência de ideias que conduziu ao resultado final.

Uma vez que o objectivo primordial do projecto é prever a inactividade de consumidores de voz móvel pré-pagos, a primeira coisa que foi feita foi definir inactividade. Primeiramente, sugeriram-se dois conceitos (apresentados e estudados paralelamente), denominados inactividade I e inactividade II.

No entanto, e como veremos na secção 2.3, a definição destes conceitos não foi de encontro ao padrão de comportamento normal de um cliente, tendo-se, por isso, posto a hipótese de definir padrões de comportamento. Uma vez estabelecido um padrão de comportamento (único para cada cliente), este poderá ser utilizado para identificar clientes cujo comportamento está em claro decréscimo.

Para tal, para cada consumidor foram definidos três limites (chamados limites de controlo inferior, *lci*, por analogia à metologia de cartas de controlo), um para a variável A, outro para a variável B e ainda outro para a variável C. Estes limites têm como base a média e o desvio padrão do cliente para cada uma destas quantidades, com base nos cinco meses anteriores ao mês em questão (neste caso, com base nos valores desde Maio a Setembro de 2012, sendo Outubro de 2012 o mês em questão).

Assumindo que serão utilizados dados para os cinco meses mais recentes, a identificação de clientes que apresentem um padrão de comportamento descendente será feita através do ajustamento de um modelo de regressão logística, que terá como resposta 1 se houver quebra de comportamento no mês seguinte ao qual o modelo está a ser aplicado ou 0 no caso contrário. Será simulada a utilização do modelo durante sete meses consecutivos (utilizando dados de Janeiro a Dezembro de 2012). A propensão de quebra de comportamento dada pelo modelo em cada um destes sete meses será categorizada, ficando disponível, para cada cliente em análise, uma sucessão de categorias de propensões de quebra. A estas sucessões será aplicada a metodologia de cadeias de Markov por forma a encontrar a categoria de risco a partir da qual um cliente já não retomará o seu padrão de comportamento habitual.

Tal como já referido anteriormente, um segundo modelo será ajustado, retirando os clientes cujos limites de controlo inferior sejam consistentemente nulos para os cinco meses em análise (por uma questão de coerência é necessário que sejam os três limites nulos durante o período em análise).

A definição formal destes conceitos, assim como a sua análise pormenorizada, será apresentada na secção *Abordagens ao problema* (secção 2.3, do capítulo 2) e a aplicação dos modelos de regressão logística, assim como a sua análise, será apresentada no capítulo *Aplicações* (capítulo 4). Também neste capítulo serão apresentados os resultados da aplicação da metodologia de cadeias de Markov ao primeiro modelo em estudo. No capítulo *Metodologia* (capítulo 3), será apresentada toda base teórica que suporta o modelo linear generalizado (e particularmente o modelo de regressão logística). Finalmente, no capítulo *Conclusões e trabalho futuro* (capítulo 5) serão apresentadas e debatidas as conclusões finais, assim como propostas para trabalho futuro.

Capítulo 2

Análise exploratória dos dados

Neste capítulo será caracterizada a população em estudo e serão apresentados os métodos de análise exploratória utilizados no desenvolvimento do projecto.

2.1 A população, a amostra e a base de dados

Uma vez contextualizado o problema, é necessário, devido ao tamanho da população em questão, retirar uma amostra representativa da mesma, por forma a ser conduzida uma análise exploratória.

Com base numa amostragem aleatória simples, foi retirada uma amostra de U consumidores que à data de 7 de Novembro de 2012 se encontravam activos. A percentagem da população representada nesta amostra é confidencial, sendo que a amostra é considerada representativa.

A base de dados da qual os dados foram recolhidos contém uma quantidade infindável de informação, constituindo, por isso, um desafio escolher quais as variáveis a analisar. Esta dissertação é baseada em dados mensais agregados (por dados agregados entende-se que se utilizam apenas contagens). Por outro lado, os dados podem estar agrupados ao nível do cliente, ao nível da conta ou ao nível do serviço. No entanto, no caso dos serviços pré-pagos, apenas interessam os dados agrupados ao nível do serviço, que é o nível mais baixo e mais simples, eliminando a preocupação com o nível segundo o qual os dados estão agrupados (cada serviço corresponde a um e um só número de telefone móvel). Assim, a partir deste ponto, um consumidor

ou cliente passará a ser genericamente designado como serviço.

De forma a não tornar maçadora a leitura desta dissertação, a análise exploratória apresentada é muitas vezes focada numa só variável, enquanto se presume igualdade de circunstâncias e conclusões para as outras variáveis em discussão. Serão referidos os casos em que tal não aconteça.

2.2 As variáveis em estudo

Nesta versão as variáveis tomarão os nomes A, B, C, D, E, F, G, H, I, J, L, M e N. As variáveis são medidas em quatro unidades distintas e apenas são comparáveis variáveis medidas na mesma unidade. A variável A é medida numa unidade, à qual se chamará m_1, as variáveis B, D, F, H e J são medidas numa unidade à qual se chamará m_2, as variáveis C, E, G, I e L são medidas numa outra unidade à qual se chamará m_3 e as restantes variáveis são medidas numa unidade à qual se chamará m_4.

Todas as variáveis estão disponíveis na base de dados para cada um dos doze meses anteriores ao mês no qual a extracção dos dados foi efectuada. Assim, se se decidir analisar os últimos seis meses de actividade um serviço, por exemplo, teremos seis valores para cada uma das variáveis referidas.

Na tabela 2.1 são apresentadas as principais medidas de localização para as variáveis estudadas, em Outubro de 2012. Para os outros meses as medidas são semelhantes. Tal como se pode verificar, os valores obtidos para as variáveis D, E, H, I, J e L são bastante baixos (geralmente 75% da amostra apresenta valores muito mais baixos para estas variáveis que para as variáveis B, C, F e G), o que provavelmente indicia que as variáveis D, E, H e I possam ser potencialmente pouco importantes, quando o objectivo é caracterizar o comportamento de um serviço.

Na figura 2.1 pode-se observar a distribuição dos valores da variável A, em m_1, para as diferentes categorias das variáveis B, C, F e G. As categorias apresentadas são diferentes para cada uma das variáveis em análise, sendo a categoria 1 definida pelo intervalo $]min, 1\ qu.]$, a categoria 2 pelo intervalo $]1\ qu., mediana]$, a categoria 3 pelo intervalo $]mediana, 3\ qu.]$ e a categoria 4 pelo intervalo $]3\ qu., max]$, em que 1 $qu.$ e 3 $qu.$ representam, respectivamente, o primeiro e o terceiro quartil. O valor dos quartís, o valor

Variável	Min.	1 Qu.	Mediana	Média	3 Qu.	Max.
A em m_1	0.00	0.00	10.00	11.75	20.00	1105.00
B em m_2	0.00	0.00	8.41	104.20	94.25	4772.0
C em m_3	0.00	0.00	1.00	294.60	121.00	23960.00
D em m_2	0.00	0.00	0.60	8.23	7.35	2828.00
E em m_3	0.00	0.00	0.00	5.09	3.00	5817.00
F em m_2	0.00	0.00	17.34	111.60	111.90	10600.00
G em m_3	0.00	0.00	7.00	284.30	128.00	16290.00
H em m_2	0.00	0.00	0.00	6.27	4.33	1422.00
I em m_3	0.00	0.00	0.00	3.89	3.00	1979.00
J em m_2	0.00	0.00	0.00	4.07	1.83	1222.00
L em m_3	0.00	0.00	0.00	1.95	1.00	2225.00
M em m_4	0.00	0.00	3.00	4.99	7.00	1455.00
N em m_4	0.00	0.00	0.00	1.73	2.00	4691.00

Tabela 2.1: Principais medidas de localização das variáveis retiradas da base de dados, em Outubro de 2012.

máximo e o valor mínimo de cada variável é como apresentado na tabela 2.1.

Como se pode verificar através da figura 2.1 o valor da variável A é mais elevado para os serviços que também apresentam valores mais elevados para cada uma das quatro variáveis representadas. No entanto, as diferenças nos valores da variável A entre as categorias 3 e as categorias 4 das variáveis apresentadas são quase nulas, muito provavelmente devido a planos de tarifários específicos aqui não considerados.

A relação entre as variáveis B, D, F, H e J e as variáveis C, E, G, I e L é similar: na generalidade, os serviços que apresentam maior número de m_2 também apresentam maior número de m_3. A mesma análise pode ser feita para as variáveis M e N, de forma a poder-se comparar o número de m_4 com o número de m_2 e m_3. Esta análise é visível na figura 2.2.

Nesta figura (2.2) mostra-se que, quanto maior o número de m_2, maior o número de m_4. A mesma conclusão pode ser retirada para as variáveis C e G.

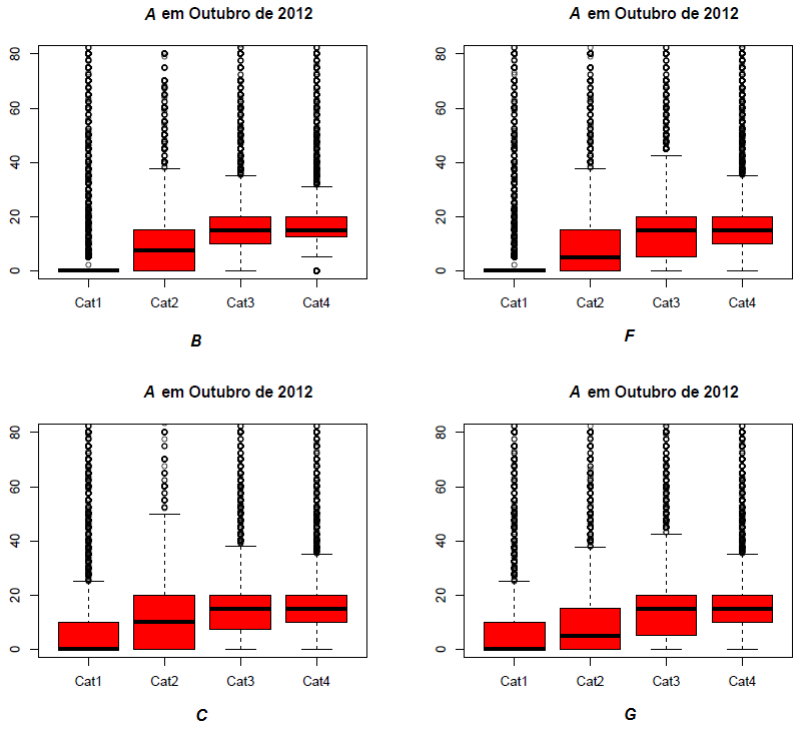

Figura 2.1: Diferenças nos valores da variável A para as diferentes categorias das variáveis B, C, F e G.

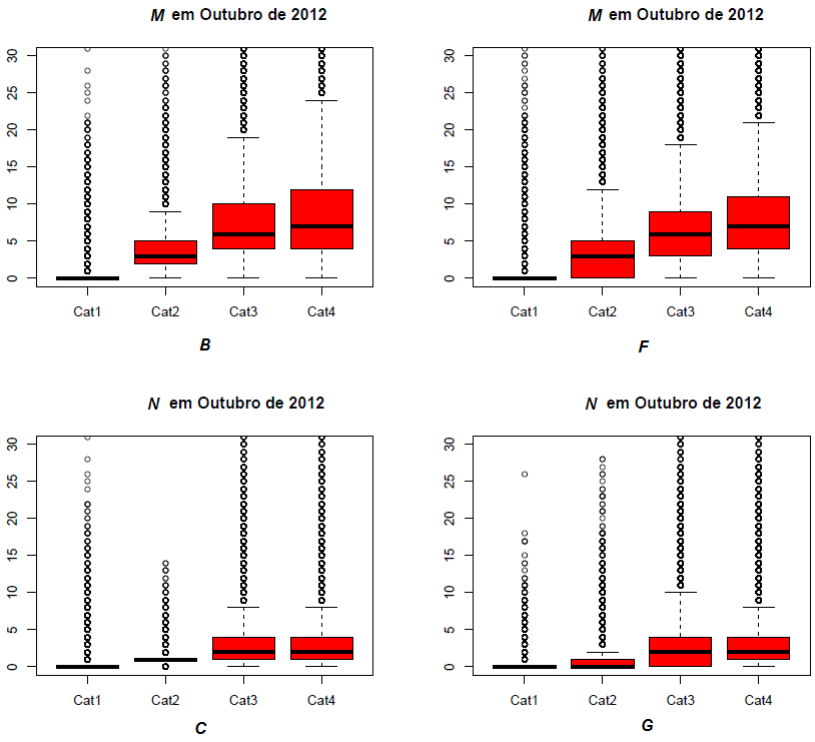

Figura 2.2: Diferenças nos valores das variáveis M e N para as diferentes categorias das variáveis B, C, F e G.

2.3 Abordagens ao problema

Tal como já foi referido no capítulo 1 a primeira abordagem ao problema consistiu em definir, paralelamente, dois conceitos distintos de inactividade, apresentados em seguida.

O conceito de inactividade I define um serviço como inactivo num certo mês quando, nesse mês, todas as variáveis até aqui referidas (excepto as variáveis M e N) tomam valor 0. Formalmente diz-se que um serviço está inactivo no mês i, segundo o conceito de inactividade I, se:

$$A = B = C = D = E = F = G = H = I = J = L = 0$$

A partir deste ponto definir-se-á um serviço como inactivo I num certo mês i se cumprir os requisitos apresentados acima, nesse mesmo mês i. Caso contrário o serviço designar-se-á como activo I.

Através da análise da figura 2.3 pode-se concluir que, desde Maio de 2012, que o valor de A para os serviços que se encontram inactivos I e para os serviços que se encontram activos I apresenta diferenças acentuadas, o que realça a importância desta variável na caracterização do estado de um serviço.

Da própria definição de inactividade I vem que o valor de M e N para os serviços que se encontram inactivos I deveriam tomar valor zero. Pode-se tomar a decisão de analisar a média destas variáveis, em vez do seu valor para um mês em concreto. É de frisar que, como sempre que se trabalha com médias, as conclusões retiradas da análise destes valores são potencialmente enganadoras, uma vez que já se sabe, à partida, que pelo menos um dos seis valores utilizados para calcular a média destas variáveis para os serviços inactivos I é zero (o valor correspondente a Outubro de 2012).

Para estes mesmos serviços, a análise relativa à média de seis meses da variável M (entre Maio e Outubro de 2012) é apresentada na tabela 2.2, enquanto que a análise relativa à média da variável N, para o mesmo período, é apresentada na tabela 2.3.

As diferenças existentes na média de M e de N de um serviço entre os serviços activos I e inactivos I em Outubro de 2012 é ainda mais visível nos histogramas apresentados na figura 2.4 e na figura 2.5, respectivamente.

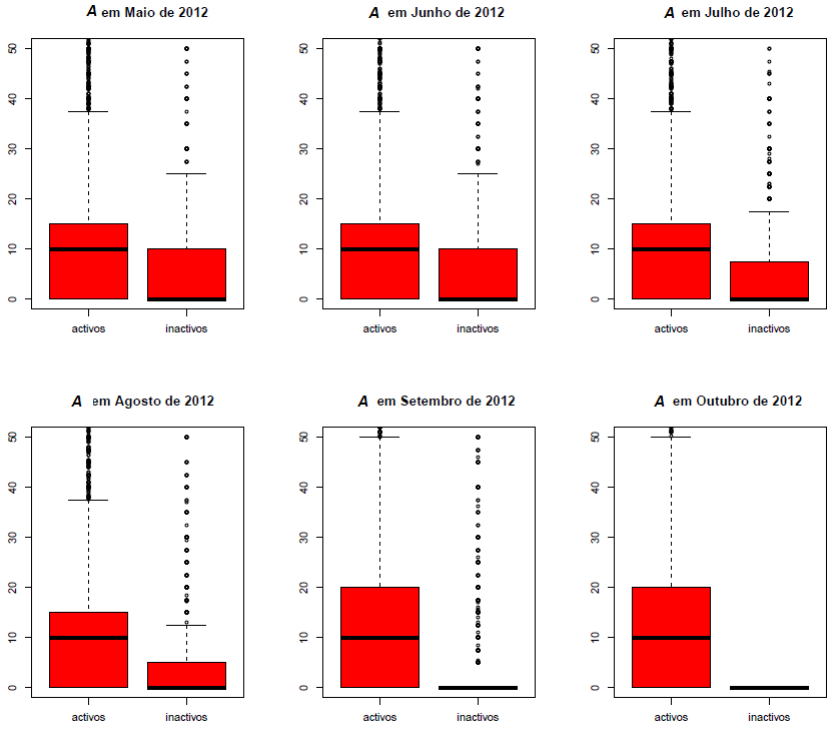

Figura 2.3: Diferenças nos valores da variável A, em m_1, entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012.

```
> round(summary(mean_M[inactivos_1]),2)
 Min.   1 Qu.   Mediana   Média   3 Qu.   Max.
0.00    0.00     0.17     0.57     0.50  178.70

> round(summary(mean_M[activos_1]),2)
 Min.   1 Qu.   Mediana   Média   3 Qu.   Max.
0.00    1.83     4.67     5.99     8.33  961.30
```

Tabela 2.2: Principais medidas de localização da variável *mean M*, para serviços inactivos I em Outubro de 2012 e para serviços activos I na mesma data, respectivamente.

```
> round(summary(mean_N[inactivos_1]),2)
 Min.     1 Qu.      Mediana      Média      3 Qu.      Max.
 0.00     0.00         0.00       0.19       0.17     63.50

> round(summary(mean_N[activos_1]),2)
 Min.     1 Qu.      Mediana      Média      3 Qu.      Max.
 0.00     0.17         1.00       2.04       2.83     781.80
```

Tabela 2.3: Principais medidas de localização da variável *mean N*, para serviços inactivos I em Outubro de 2012 e para serviços activos I na mesma data, respectivamente.

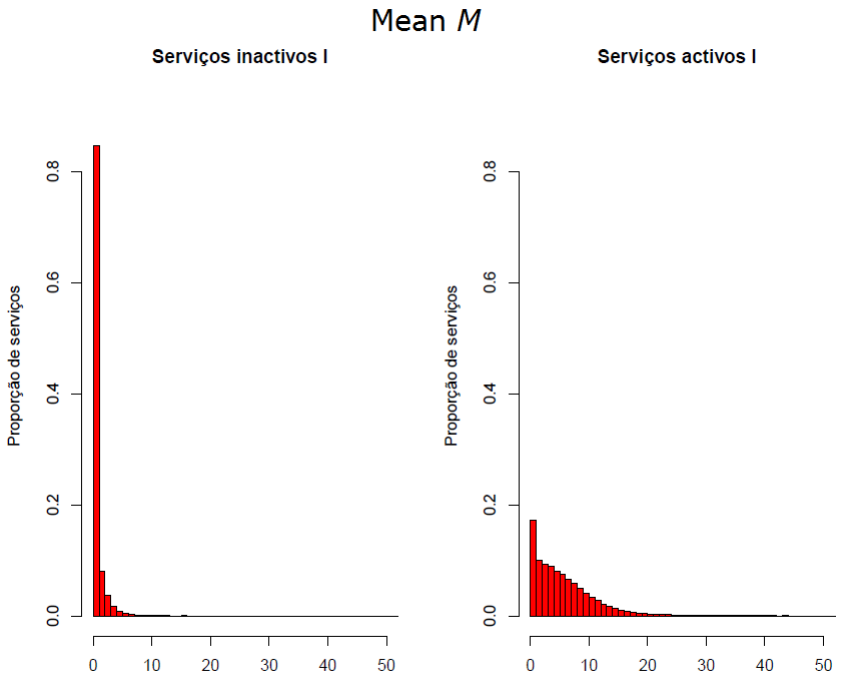

Figura 2.4: Diferenças da média dos valores da variável M entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012.

Nos histogramas apresentados na figura 2.4 é clara a diferença entre o número médio de M, em m_4, de um serviço que se encontrava no estado inactivo I em Outubro de 2012 e um serviço que se encontrava no estado activo I no mesmo mês.

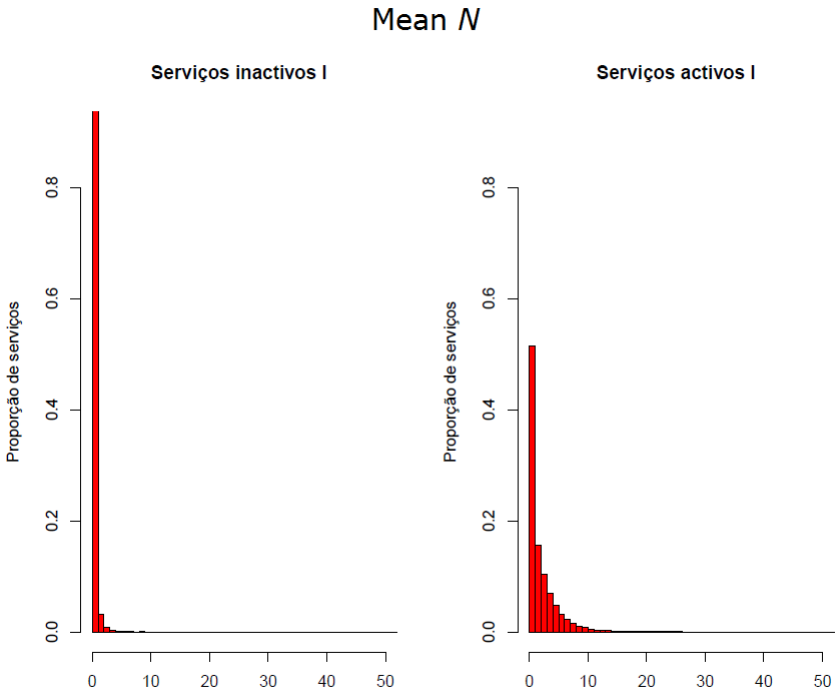

Figura 2.5: Diferenças da média dos valores da variável N entre os serviços que se encontravam inactivos I e os serviços que se encontravam activos I, em Outubro de 2012.

Mais uma vez, e através da análise dos histogramas apresentados na figura 2.5, é clara a diferença entre o número médio de N, em m_4, de um serviço que se encontrava no estado inactivo I em Outubro de 2012 e um serviço que se encontrava no estado activo I no mesmo mês.

Tal como referido, o conceito de inactividade I foi desenvolvido em paralelo com um outro conceito de inactividade, chamado inactividade II. O conceito de inactividade II é menos restritivo e define um serviço como inac-

tivo num certo mês quando, nesse mês, as variáveis A, B, C, D e E tomam valor nulo. Formalmente diz-se que um serviço está inactivo no mês i, segundo o conceito de inactividade II, se:

$$A = B = C = D = E = 0$$

A partir deste ponto definir-se-á um serviço como inactivo II num certo mês i se cumprir os requisitos apresentados acima, nesse mesmo mês i. Caso contrário o serviço designar-se-á como activo II.

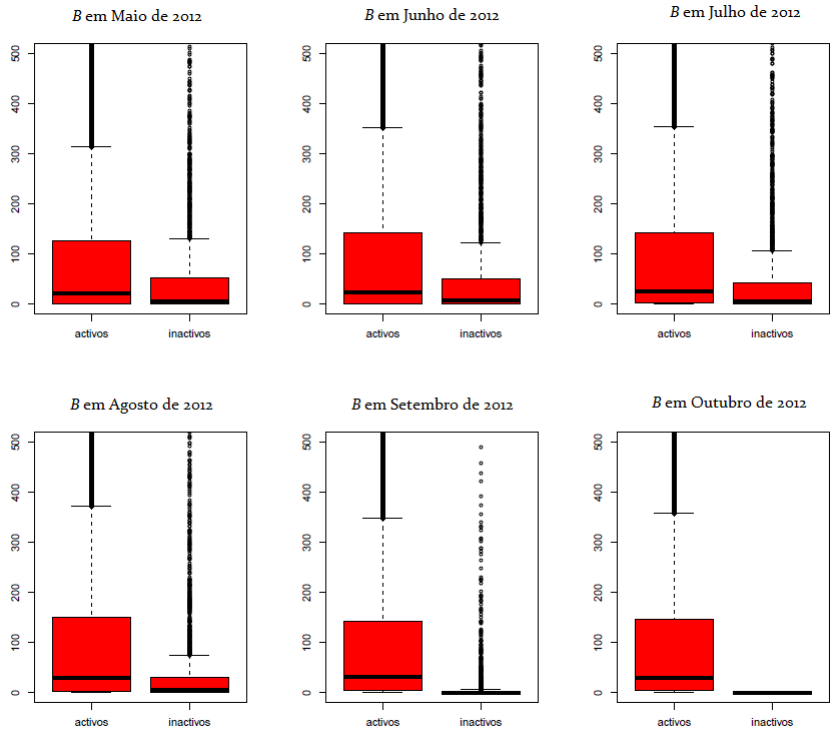

Figura 2.6: Diferenças nos valores da variável B entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012.

Através da análise da figura 2.6 pode-se concluir que, desde Maio de 2012, que o valor de B para os serviços que se encontram inactivos II e para os serviços que se encontram activos II apresenta diferenças acentuadas, o que realça a importância desta variável na caracterização do estado de um

```
> round(summary(M_Out_12[inactive_2]),2)
 Min.     1 Qu.      Mediana      Média      3 Qu.      Max.
0.00      0.00        0.00        0.03       0.00      25.00

> round(summary(M_Out_12[active_2]),2)
 Min.     1 Qu.      Mediana      Média     3 Qu.      Max.
 0.00     2.00         5.00       6.46      9.00     1455.00
```

Tabela 2.4: Principais medidas de localização da variável M, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente.

```
> round(summary(N_Out_12[inactive_2]),2)
 Min.     1 Qu.      Mediana      Média      3 Qu.      Max.
0.00      0.00        0.00        0.02       0.00      26.00

> round(summary(N_Out_12[active_2]),2)
 Min.     1 Qu.      Mediana      Média     3 Qu.      Max.
0.00      0.00        1.00        2.24      3.00     4691.00
```

Tabela 2.5: Principais medidas de localização da variável N, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente.

serviço.

Ao contrário da inactividade I, a inactividade II não pressupõe que as variáveis M e N tomem valor zero, podendo-se, portanto, analisar o valor exacto das variáveis para o mês de Outubro de 2012 (ou para outro mês qualquer). As principais medidas de localização para estas variáveis no período já referido são apresentadas na tabela 2.4 e na tabela 2.5.

Para os serviços em questão, a análise relativa à média de seis meses da variável M (entre Maio e Outubro de 2012) é apresentada na tabela 2.6, enquanto que a análise relativa à média da variável N, para o mesmo período, é apresentada na tabela 2.7.

Mais uma vez, as diferenças existentes na média de M e de N entre os serviços activos II e inactivos II em Outubro de 2012 é ainda mais visível nos histogramas apresentados na figura 2.7 e na figura 2.8, respectivamente.

Nos histogramas apresentados na figura 2.7 é clara a diferença entre o número médio de M, em m_4, de um serviço que se encontrava no estado inactivo II em Outubro de 2012 e um serviço que se encontrava no estado

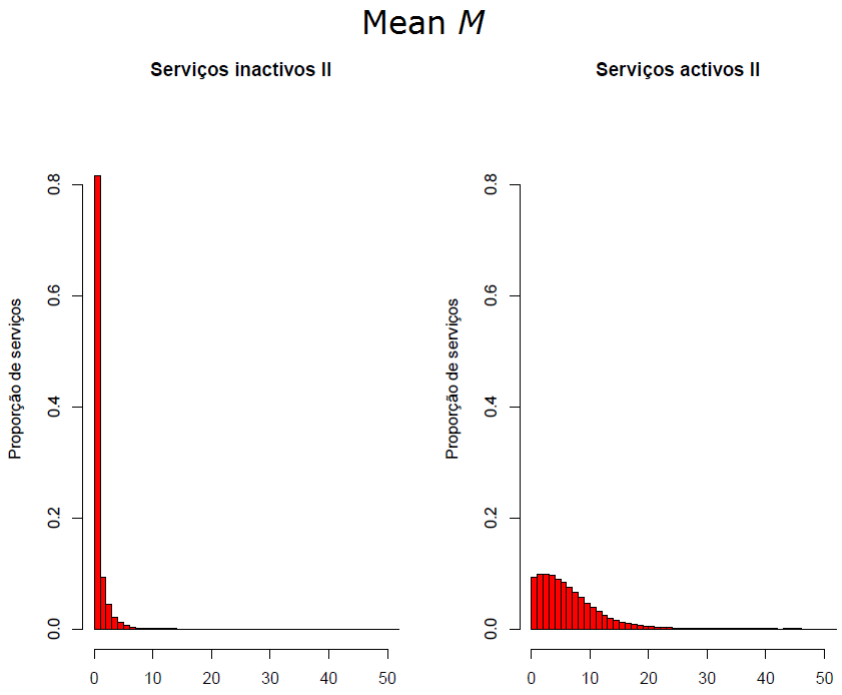

Figura 2.7: Diferenças da média dos valores da variável *mean M* entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012.

```
> round(summary(mean_M[inactive_2]),2)
 Min.    1 Qu.     Mediana     Média     3 Qu.     Max.
0.00     0.00       0.17       0.66      0.67     178.70

> round(summary(mean_M[active_2]),2)
 Min.    1 Qu.     Mediana     Média    3 Qu.     Max.
0.00     2.67       5.33       6.69     9.00    961.30
```

Tabela 2.6: Principais medidas de localização da variável *mean M*, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente.

```
> round(summary(mean_N[inactive_2]),2)
 Min.    1 Qu.     Mediana     Média     3 Qu.     Max.
0.00     0.00         0.00       0.26     0.17     64.50

> round(summary(mean_N[active_2]),2)
 Min.    1 Qu.     Mediana     Média    3 Qu.     Max.
0.00     0.33       1.33       2.27     3.17     781.80
```

Tabela 2.7: Principais medidas de localização da variável *mean N*, para serviços inactivos II em Outubro de 2012 e para serviços activos II na mesma data, respectivamente.

activo II no mesmo mês.

Mais uma vez, e através da análise dos histogramas apresentados na figura 2.8, é clara a diferença entre o número médio de N, em m_4, de um serviço que se encontrava no estado inactivo II em Outubro de 2012 e um serviço que se encontrava no estado activo II no mesmo mês.

De certa forma o conceito de inactividade I está contido no conceito de inactividade II, sendo que um serviço que seja inactivo I também é inactivo II (embora o reciproco não seja verdade). Na realidade, espera-se que um serviço passe de um estado de actividade, para um estado de inactividade II e só depois para um estado de inactividade I.

Sendo o conceito de inactividade II menos restritivo que o de inactividade I espera-se que haja mais serviços num estado de inactividade II do que num estado de inactividade I. Tal já foi confirmado, uma vez que, em Outubro de 2012, há mais 10% de serviços inactivos II do que inactivos I. Dado o tamanho da população, isto pode ser (e é) um problema. No entanto,

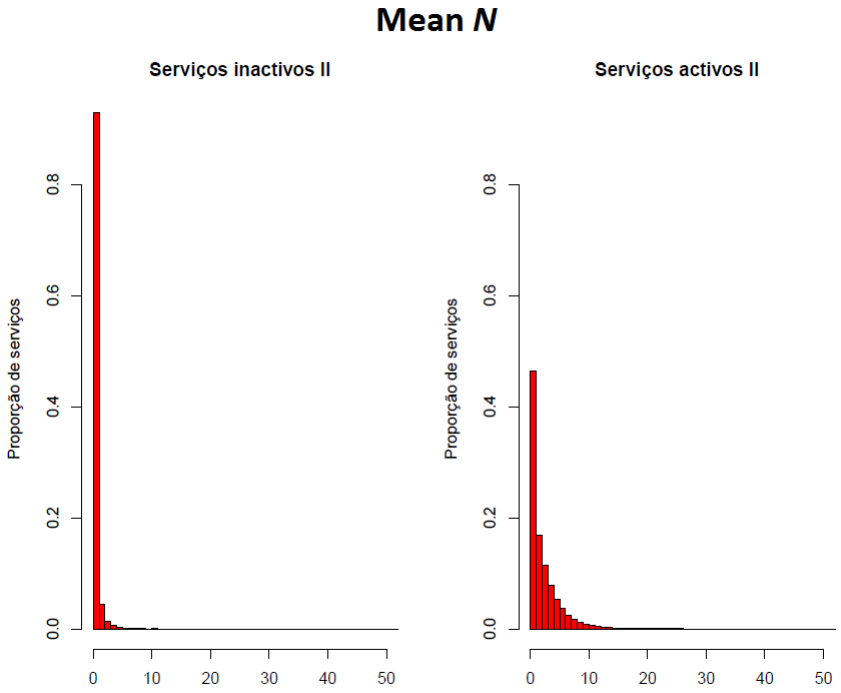

Figura 2.8: Diferenças da média dos valores da variável *mean N* entre os serviços que se encontravam inactivos II e os serviços que se encontravam activos II, em Outubro de 2012.

e ponderando os pros e os contras, é de maior interesse da empresa detectar um serviço que passe de um estado de actividade para um estado de inactividade II do que detectar um serviço que passe de um estado de actividade para um estado de inactividade I, uma vez que poderá ser demasiado tarde para recuperar o comportamento do serviço.

No entanto, a formulação destes dois conceitos tem como pilar a suposição de que um serviço está constantemente activo, sendo o objectivo prever a primeira vez que esse mesmo serviço entra num estado de inactividade (seja I ou II). De forma a verificar se este pressuposto está bem ajustado à amostra em estudo (e consequentemente à população), da amostra de U serviços iniciais retiraram-se os serviços cuja data de activação era posterior a Maio de 2012, isto é, serviços para os quais não existia um histórico de pelo menos seis meses.

Para cada um dos dois conceitos de inactividade introduzidos anteriormente, estes serviços foram divididos em dois grupos: o grupo dos serviços que se encontravam activos em Outubro de 2012 e o grupo dos que se encontravam inactivos, no mesmo período. De seguida, para cada serviço, foi calculado o estado de actividade em que se encontrava no mês de Junho, de Julho, de Agosto e de Setembro de 2012, segundo os dois critérios de actividade considerados, e agruparam-se os serviços com sequências de estados de actividade iguais.

Com base na análise efectuada pode-se, então, concluir que não se verifica o pressuposto de que um serviço será, maioritariamente, activo e apenas raramente inactivo. De facto, qualquer que seja o conceito de inactividade utilizado, a percentagem de serviços que alternam entre estados de actividade e inactividade é bastante significativa, pondo em causa o rumo seguido até este ponto, uma vez que prever inactividade para um serviço num certo mês pode não ser relevante pois tal comportamento pode estar de acordo com o seu padrão de comportamento.

Assim, põe-se a questão de até que ponto não seria mais vantajoso para a empresa definir-se um padrão de comportamento para cada serviço, baseado no seu comportamento para os últimos meses, permitindo identificar os consumidores cujo comportamento está num padrão claramente descendente.

A primeira ideia em relação a esta abordagem consistiu na criação de categorias para as variáveis A, B e C, únicas para cada serviço, baseadas na

média e desvio-padrão do mesmo, para cada uma destas variáveis.

Neste ponto é importante frisar que apenas se consideraram estas variáveis uma vez que durante a análise exploratória, foi encontrada evidência de que as variáveis D, E, F, G, H, I, J, L, M e N não serão relevantes para a caracterização do comportamento, o que as exclui deste estudo.

Para a criação destas categorias, tanto a média como o desvio padrão foram calculados utilizando os valores para cada uma das três variáveis referidas desde Maio a Setembro de 2012. A ideia inicial seria existirem cinco categorias para cada uma destas três variáveis, definidas como na figura 2.9, no caso em que a quantidade média menos desvio padrão é positiva ou quatro categorias, como definidas na figura 2.10, no caso em que a quantidade média menos desvio padrão é negativa.

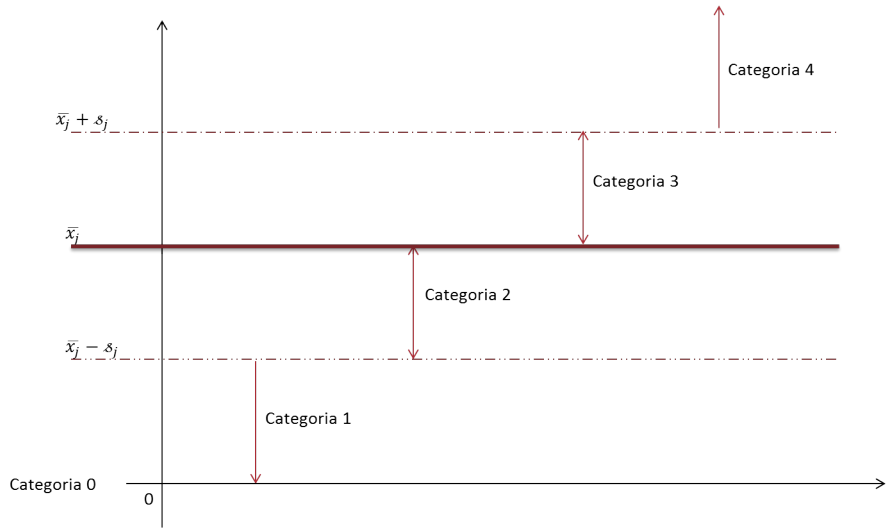

Figura 2.9: Categorias de comportamento para um serviço j, no caso em que $x_j - s_j > 0$.

O objectivo desta abordagem seria utilizar um modelo de regressão logística multinomial para prever em que categoria das variáveis A, B e C é que um serviço se iria encontrar no mês seguinte ao mês em que o modelo seria aplicado. Um modelo de regressão logística multinomial é uma generalização do modelo de regressão logística ao permitir mais do que dois desfechos discretos, isto é, é um modelo que é usado para prever as probabilidades

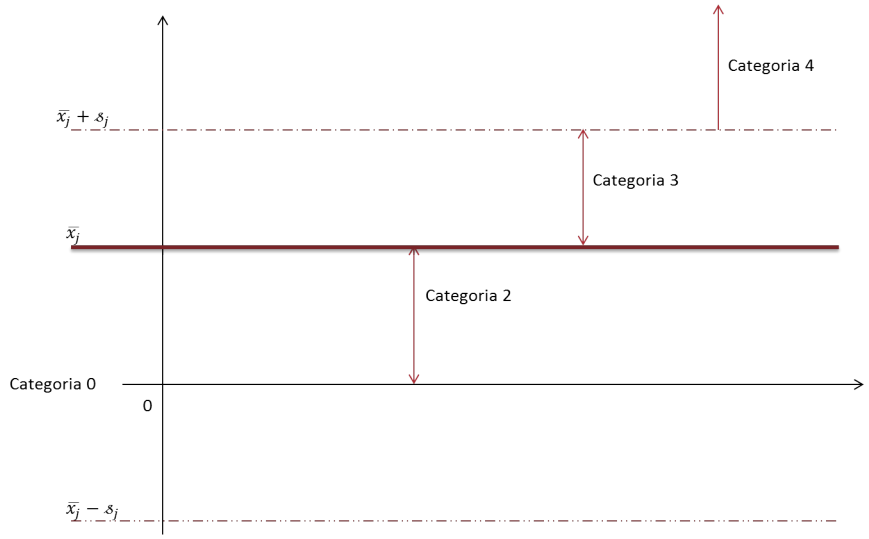

Figura 2.10: Categorias de comportamento para um serviço j, no caso em que $x_j - s_j \leq 0$.

dos diferentes desfechos possíveis de uma variável resposta categoricamente distribuída, dado um conjunto de variáveis independentes (que podem ser contínuas, binárias, categóricas, etc..).

Apesar de não haver muito sentido prático em ajustar três modelos distintos para prever as categorias de três variáveis distintas, a principal desvantagem desta abordagem é o facto de não se saber como conjugar os valores das três variáveis de forma a se ter uma informação única sobre o comportamento do serviço, isto é, se um serviço estiver, por exemplo, na categoria 1 na variável A e na categoria 4 nas variáveis B e C como é que esta informação se reflecte em termos de comportamento? É necessário que o serviço seja alvo de uma campanha para estimular o comportamento ou o simples facto de apresentar duas categorias elevadas exclui o mesmo como potencial candidato a uma acção de retenção?

De certa forma, o que se pretende é encontrar a combinação óptima que traduza num único valor a informação relevante relativa às três variáveis em estudo. Assim sendo, não é necessário utilizar as categorias de comportamento tais como definidas nas figuras 2.9 e 2.10, uma vez que se poderão usar os valores concretos das variáveis, eliminando a inevitável perda de informação decorrente da categorização das variáveis.

	A	B	C
A	1	0.22	0.09
B	0.22	1	0.33
C	0.09	0.33	1

Tabela 2.8: Correlação entre as variáveis A, B e C.

A melhor forma de encontrar a combinação linear que explica a maior percentagem da variabilidade das variáveis envolvidas é utilizar a análise em componentes principais. De forma simplista, a análise em componentes principais transforma um conjunto de observações de variáveis possivelmente correlacionadas num conjunto de valores de variáveis linearmente não correlacionadas, chamadas componentes principais. O número de componentes principais é menor ou igual ao número de variáveis envolvidas e a primeira componente principal é sempre a que explica maior proporção da variância das variáveis originais. A análise em componentes principais é sensível à escala na qual as variáveis originais são medidas, no sentido em que as variáveis devem ser todas medidas na mesma escala (neste caso as três variáveis são medidas em escalas distintas, a variável A é medida em m_1, a variável B é medida em m_2 e a variável C é medida em m_3). No entanto, este facto é anulado se se usar a matriz de correlações em vez da matriz de covariâncias das observações para calcular as componentes principais.

Tal como referido, um dos principais requisitos para que esta metodologia seja aplicada é que as variáveis originais sejam correlacionadas, sendo que, quanto mais correlacionadas forem, mais ganho provém da análise em componentes principais. Neste caso, tal como é apresentado na tabela 2.8, a correlação entre as variáveis é baixa (apesar de não ser nula).

Assim sendo, a análise em componentes principais não dará resultados satisfatórios, uma vez que sendo as variáveis pouco correlacionadas, é praticamente impossível encontrar um valor que resuma satisfatoriamente as três. No entanto persiste o problema de como combinar os valores destas três variáveis de forma a que o comportamento de um serviço esteja bem explicado.

Voltando ao conceito das categorias de comportamento (e uma vez que os modelos de regressão logística multinomial, quando ajustados, revelaram fraco desempenho ao predizer categorias futuras), foram definidos limites de

controlo inferior (lci), únicos para cada serviço j, para o valor de A em m_1, para o valor de B, em m_2, e para o valor de C, em m_3, como se apresentam em seguida:

$$lci\ A_j = \overline{A_j} - 2 * s(A_j) \tag{2.1}$$

$$lci\ B_j = \overline{B_j} - 2 * s(B_j) \tag{2.2}$$

$$lci\ C_j = \overline{C_j} - 2 * s(C_j) \tag{2.3}$$

As médias e desvios padrões apresentados são calculados com os dados relativos aos últimos cinco meses disponíveis, neste caso de Maio a Setembro de 2012.

Estes limites de controlo inferior representam o limiar a partir do qual a variação no comportamento é considerada maior do que o usual, sendo que abaixo deste valor considera-se que o consumidor entrou num padrão de comportamento claramente descendente.

No entanto, e antes de se prosseguir, é necessário verificar quantos são os serviços que apresentam valor zero para o conjunto dos três limites, uma vez que isto se pode dever a uma de duas situações:

1. Para cada uma destas quantidades a média do serviço é igual a duas vezes o seu desvio padrão
2. O valor de cada uma destas quantidades é 0 há pelo menos cinco meses, o que retorna uma média e um desvio padrão nulos

Todos os serviços que satisfazem estas condições encontram-se na opção 2. Aquando do ajuste dos modelos de regressão logística ambas as amostras são consideradas: a amostra completa e a amostra sem os serviços que apresentam os três limites nulos.

Existem serviços cuja variabilidade é muito elevada, fazendo com que os limites inferiores definidos nas equações (2.1), (2.2) e em (2.3) sejam negativos, tal como se pode ver na tabela 2.9. Assim é atribuído o valor 0 quando

```
> round(summary(lci_A),2)
  Min.     1 Qu.     Mediana     Média    3 Qu.     Max.
-414.40    -5.21       0.00      -0.68    4.58     236.40

> round(summary(lci_B),2)
  Min.     1 Qu.     Mediana    Média    3 Qu.     Max.
-2728.00   -8.59      -0.50     5.18     5.41     4224.00

> round(summary(lci_C),2)
  Min.     1 Qu.     Mediana     Média     3 Qu.     Max.
-9804.00   -5.55      0.00       45.83     0.13   13950.00
```

Tabela 2.9: Principais medidas de localização os limites de controlo inferior para as variáveis A, B e C, respectivamente.

o limite inferior é negativo, uma vez que é necessário um valor não negativo com o qual comparar os valores de A, B e C.

Uma vez estabelecidos os limites a partir dos quais se crê que um serviço se encontra num padrão claramente descendente, é necessário definir um critério que permita decidir se um serviço vai sofrer quebra no seu padrão de comportamento ou não, uma vez que existem três limites, e que um limite ser ultrapassado num certo mês não implica que os outros também sejam.

Para tal criou-se um conjunto de variáveis binárias (denominadas variáveis *down*) que, para cada mês i e para cada serviço j, tomam valor 1 caso, no mês i o valor de A, de B ou de C seja menor que o limite de controlo inferior respectivo. Caso contrário tomam valor 0. Formalmente, este conjunto de três variáveis para cada mês i e para cada serviço j, são definidas nas equações (2.4), (2.5) e (2.6).

$$down\ A_j^i = \begin{cases} 0, & \text{se } A_j^i > lci\ A_j \\ 1, & \text{se } A_j^i \leq lci\ A_j \end{cases} \tag{2.4}$$

$$down\ B_j^i = \begin{cases} 0, & \text{se } B_j^i > lci\ B_j \\ 1, & \text{se } B_j^i \leq lci\ B_j \end{cases} \tag{2.5}$$

$$down\ C_j^i = \begin{cases} 0, & \text{se } C_j^i > lci\ C_j \\ 1, & \text{se } C_j^i \leq lci\ C_j \end{cases} \tag{2.6}$$

Em relação às variáveis *down* é de frisar que o limite de controlo inferior com

o qual o valor das variáveis é comparado é calculado com base nos dados disponíveis para os cinco meses mais recentes. Assim, por exemplo, o valor da variável *down* A_j^{Maio} é comparado com o *lci* A_j calculado em Outubro de 2012, com base nos valores para a variável A desde Maio a Setembro de 2012.

Mais uma vez, o problema de como conjugar os valores de A, de B e de C, para cada serviço, mantém-se.

Para tal foi definido que o serviço j sofre quebra no padrão comportamento no mês i se o valor de B e de C estiver abaixo dos respectivos limites de controlo inferior, para o serviço j no mês i, isto é, se o valor das variáveis *down* B e *down* C for 1. Formalmente, a variável que indica se no mês i houve quebra no padrão de comportamento para o serviço j, chamada $break_j^i$, é uma variável binária, que toma valor 1 caso se cumpram as condições explicitadas na equação (2.7).

$$\begin{cases} B_j^i \leq lci\ B_j \\ C_j^i \leq lci\ C_j \end{cases} \tag{2.7}$$

As condições especificadas na equação (2.7) são equivalentes às condições especificadas na equação (2.8).

$$\begin{cases} down\ B_j^i = 1 \\ down\ C_j^i = 1 \end{cases} \tag{2.8}$$

Obviamente que caso ambas as condições não sejam cumpridas, a variável $break_j^i$ toma valor 0.

É de frisar que na definição das variáveis *break* não foi utilizada informação relativa à variável A uma vez que a quebra no valor desta variável, quando considerada individualmente, não se revelou representativa da quebra de comportamento.

Para ajustar o modelo de regressão logística apresentado no capítulo 4, as variáveis *down* serão utilizadas como variáveis independentes, enquanto a variável *break*, correspondente ao mês i que se quer prever, será a variável resposta.

Capítulo 3

O modelo linear generalizado

Um modelo é uma abstracção da realidade uma vez que fornece uma aproximação de um qualquer fenómeno relativamente mais complexo (Myers (2002)). Os modelos existentes podem ser, rudemente, classificados como determinísticos ou probabilístico. No caso dos modelos probabilísticos, a resposta dada pelo modelo exibe variabilidade, porque o modelo ou contém elementos aleatórios ou sofre o impacto de forças aleatórias. A classe de modelos probabilísticos mais importante é a classe dos modelos lineares:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \epsilon, \tag{3.1}$$

onde Y é a variável resposta, $X_1, X_2, ..., X_k$ é um conjunto de variáveis regresssoras (ou preditoras, ou covariáveis), $\beta_0, \beta_1, ..., \beta_k$ é um conjunto de parâmetros de regressão desconhecidos e ϵ é o erro aleatório. À equação (3.1) dá-se o nome de modelo linear. Um dos principais pressupostos do modelo linear é que a variável resposta Y segue uma distribuição normal com valor médio dado pela equação (3.2).

$$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3.2}$$

O modelo linear generalizado é, tal como o nome indica, uma generalização do modelo linear para casos em que a variável resposta segue uma distribuição pertencente à familia exponencial, correspondendo à relaxação da condição imposta pelo modelo linear de que a variável resposta deve seguir

uma distribuição Normal.

Diz-se que uma variável aleatória (v.a.) X tem distribuição pertencente à familia exponencial se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma

$$f(x|\theta,\phi) = exp\left\{\frac{x\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.3}$$

onde θ é a forma canónica do parâmetro de localização, ϕ é um parâmetro de dispersão suposto conhecido e $a(\cdot)$, $b(\cdot)$ e $c(\cdot,\cdot)$ são funções reais conhecidas. Admite-se ainda que $b(\cdot)$ é diferenciável e que o suporte da distribuição não depende dos parâmetros (Turkman (2000)).

Mais, se X é uma v.a. com distribuição pertencente à família exponencial, tal como definida em (3.3), tem-se:

1. $E(X) = \mu = b'(\theta)$
2. $Var(X) = a(\phi)b''(\theta)$

A familia exponencial inclui distribuições discretas e continuas tais como a distribuição Normal, a distribuição Binomial, a distribuição Geométrica, a distribuição de Poisson, a distribuição Binomial Negativa, a distribuição Exponencial e a distribuição Gama.

Tomando como exemplo o caso de uma v.a. T com distribuição Binomial com parâmetros n e p, $T \sim Bin(n,p)$ e com f.m.p. dada por:

$$\mathcal{P}(T=t) = \binom{n}{t} \cdot p^t \cdot (1-p)^{n-t} \tag{3.4}$$

$$= exp\left\{log\binom{n}{t} + t \cdot log(p) + (n-t) \cdot log(1-p)\right\} \tag{3.5}$$

$$= exp\left\{log\binom{n}{t} + t \cdot log\left(\frac{p}{1-p}\right) + n \cdot log(1-p)\right\} \tag{3.6}$$

$$= exp\left\{t \cdot \theta - n \cdot log\left(1+e^{\theta}\right) + log\binom{n}{t}\right\}, \tag{3.7}$$

com $\theta = log\left(\frac{p}{1-p}\right)$ e $t = 0, 1, 2, 3, ...$

Vê-se assim que esta f.m.p é da forma (3.3) com:

- $\theta = log\left(\frac{p}{1-p}\right)$ - função *logit*;
- $b(\theta) = n \cdot log\left(1 + e^{\theta}\right)$;
- $c(y, \phi) = log\binom{n}{t}$;
- $a(\phi) = 1$;
- $b'(\theta) = n \cdot \frac{e^{\theta}}{1+e^{\theta}} = n \cdot p$;
- $a(\phi) \cdot b''(\theta) = 1 \cdot n \cdot \frac{e^{\theta}}{1+e^{\theta}} \cdot \frac{1}{1+e^{\theta}} = n \cdot p \cdot (1-p)$.

A extensão mencionada anteriormente é feita em duas direcções: além da já referida relaxação da condição de que a variável resposta Y siga uma distribuição Normal, podendo seguir qualquer distribuição pertencente à familia exponencial, a função que relaciona o valor esperado e o vector de covariáveis (chamada função de ligação e representada por $g(\cdot)$) pode ser qualquer função diferenciável. Formalmente, o modelo linear generalizado pode ser definido como na equação (3.8) (Myers (2002)).

$$g(\mu) = g[E(Y)] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k \tag{3.8}$$

Qualquer modelo linear generalizado é constituído por 3 partes:

- uma distribuição para a variável resposta Y;
- um preditor linear que envolva as variáveis regressoras;
- uma função de ligação $g(\cdot)$ que liga o preditor linear ao valor médio da variável resposta.

Quando a função de ligação entre o valor médio da variável resposta e o preditor linear é a função *logit* (correspondendo a uma variável resposta com distribuição Binomial) o modelo linear generalizado toma o nome particular de modelo de regressão logística,ou modelo *logit*.

3.1 O modelo de regressão logística

Em estatística, a regressão logística é um tipo de análise de regressão usada para predizer o resultado de uma variável dependente categórica, baseada numa ou mais variáveis preditoras. As probabilidades dos possíveis desfechos são modeladas como uma função das variáveis explicativas, usando a função logística. A regressão logística pode ser binomial ou multinomial (tal como já foi referido na secção 2.3). No caso da regressão logística binomial, a variável resposta toma dois valores distintos, normalmente 0 e 1, sendo 1 a codificação atribuída ao sucesso, cuja ocorrência se pretende prever.

A regressão logística é utilizada para predizer os *odds* da resposta $Y = 1$ baseado nas variáveis preditivas. Os *odds* de $Y = 1$ são definidos como a probabilidade da variável resposta tomar o valor 1 dividida pela probabilidade de tomar o valor 0. Formalmente:

$$Odds = \frac{\mathcal{P}(Y = 1)}{\mathcal{P}(Y = 0)} = \frac{p}{1 - p} \tag{3.9}$$

Tomando como exemplo ilustrativo o caso em que apenas existe uma covariável, de forma a se perceber como é que a função logística é utilizada como função de ligação, tem-se:

$$p(x) = \frac{e^{(\beta_0 + \beta_1 x)}}{e^{(\beta_0 + \beta_1 x)} + 1} = \frac{1}{e^{-(\beta_0 + \beta_1 x)} + 1}, \tag{3.10}$$

$$g(x) = log \frac{p(x)}{1 - p(x)} = \beta_0 + \beta_1 x \tag{3.11}$$

e

$$\frac{p(x)}{1 - p(x)} = e^{(\beta_0 + \beta_1 x)} \tag{3.12}$$

O *input* é $\beta_0 + \beta_1 x$ e o *output* é $p(x)$. A função logística é útil porque pode tomar como *input* qualquer valor pertencente ao intervalo $(-\infty, +\infty)$, enquanto que o *output* é confinado ao intervalo $[0, 1]$. Nas equações (3.10) - (3.12), $g(x)$ refere-se à função logística de um certo preditor conhecido X, $p(x)$ é a probabilidade de $Y = 1$ desconhecida, β_0 é o *intercept* da equação linear (é o valor da equação quando o preditor toma valor zero) e $\beta_1 x$ é o coeficiente da regressão multiplicado por algum valor x do preditor. A

primeira equação (equação (3.10)) mostra que a probabilidade de $Y = 1$ é igual à função logística da equação de regressão linear. Isto é importante no sentido que implica que o *input* da equação da regressão logística (a equação da regressão linear) pode variar no intervalo $(-\infty, +\infty)$ e que, ainda assim, após exponenciar os *odds* da equação, o *output* pode variar no intervalo $[0, 1]$. A segunda equação (equação (3.11)) ilustra que o *logit* (também chamado *log-odds* ou logaritmo natural dos *odds*) é equivalente à equação de regressão linear. Da mesma forma, a terceira equação (equação (3.12)) ilustra que os *odds* de $Y = 1$ são equivalentes à função exponencial da equação de regressão linear. Assim, é demonstrada a importância de utilizar o *logit* função de ligação entre a probabilidade e a equação de regressão linear pois, além de variar no intervalo $[0, 1]$, é facilmente convertido nos *odds*.

3.1.1 Ajustamento do modelo

Os coeficientes da regressão são normalmente estimados usando o método de máxima verosimilhança. Ao contrário da regressão linear, na qual os resíduos seguem uma distribuição Normal, não é possível encontrar uma expressão fechada para o valor dos coeficientes que maximiza a função de verosimilhança, tendo, então, que se usar um processo iterativo, como o método de Newton, por exemplo. Tal como a maioria dos processos iterativos, este começa com uma solução inicial, altera-a ligeiramente por forma a perceber se pode ser melhorada e repete este processo até que o melhoramento seja ínfimo. Quando tal acontece diz-se que o processo converge.

Há casos nos quais o processo não atinge a convergência. Nestes casos, os coeficientes da regressão não são significativos porque o processo iterativo não foi capaz de fornecer uma solução apropriada. Isto pode acontecer por variadas razões: multicolinearidade, separação completa ou dispersão, por exemplo.

Embora não seja um valor preciso, os modelos de regressão logística requerem um mínimo de dez eventos por variável explicativa (onde eventos denota os casos pertencentes à categoria menos frequente da variável dependente). Ter uma grande proporção de sucessos resulta numa estatística de Wald ultra conservadora (abordado em pormenor mais à frente) e pode levar à não convergência do processo iterativo.

O conceito de multicolinearidade refere-se a uma correlação inaceita-

velmente alta entre preditores. Quando a multicolinearidade aumenta, os coeficientes mantêm-se centrados mas os erros padrões aumentam e a verosimilhança do modelo diminui. Para detectar multicolinearidade entre os preditores pode-se conduzir uma análise de regressão linear com os preditores de interesse, tendo como único propósito examinar a tolerância estatística usada para decidir se a multicolinearidade é inaceitavelmente alta (Menard (1995)), em alternativa à abordagem *standard* que consiste em calcular o coeficiente de correlação de Pearson para avaliar a associação entre preditores. Outra forma de detectar a multicolinearidade entre os preditores é calculando o *variance inflation factor*. Em estatística, o *variance inflation factor* (VIF) quantifica a importância da multicolinearidade numa análise de regressão. Este factor fornece um índice que mede quanto é que a variância de um coeficiente da regressão aumenta devido à colinearidade. Está convencionado que quando um valor de VIF é maior do que 5 pode-se afirmar que a multicolinearidade é elevada (apesar de haver alguns autores que tomam o valor 10 como valor fronteira (Kutner (2004))).

Tal como na análise de regressão linear, na qual é utilizada a soma de quadrados, é necessário avaliar a qualidade de ajustamento do modelo. No caso do modelo de regressão logística essa medida toma o nome de *deviance*. A *deviance* é análoga à soma de quadrados do modelo linear e é uma medida dos desvios no ajuste de um modelo de regressão logística aos dados em causa. É calculada comparando o modelo corrente (modelo em análise) com o modelo saturado (modelo no qual o número de variáveis é igual ao número de observações), sendo que à computação do valor da *deviance* também se dá o nome de teste de razão de verosimilhanças. A *deviance* é dada por (Hosmer (2000)):

$$Deviance = -2 \cdot log\left(\frac{verosimilhança\ do\ modelo\ corrente}{verosimilhança\ do\ modelo\ saturado}\right) \tag{3.13}$$

A razão entre a verosimilhança do modelo corrente e a verosimilhança do modelo saturado é sempre um número negativo, sendo multiplicado pelo escalar -2 por forma a produzir valores com uma distribuição aproximadamente χ^2. Valores pequenos da *deviance* indicam melhores ajustamentos uma vez que o modelo corrente desvia-se menos do modelo saturado. Quando avaliados segundo uma distribuição de χ^2, valores não significativos indicam muito pouca variância por explicar e, consequentemente, um bom ajustamento. Contrariamente, valores significativos segundo uma distribuição de

χ^2 indicam uma quantidade representativa de variância por explicar.

Além do modelo corrente e do modelo saturado existe também o modelo nulo, que tem um único parâmetro representativo para todos os $Y_i's$, e que entende que toda a variação nos dados é devida à componente aleatória.

Outra forma de avaliar o ajustamento do modelo corrente é utilizando o critério de informação de Akaike (AIC) que avalia a log-verosimilhança do modelo, penalizando esse valor com o número de covariáveis que entram no modelo. Um modelo com menos preditores terá um AIC mais baixo e será, portanto, positivamente avaliado. Este critério oferece uma medida relativa da informação perdida por um determinado modelo sendo que quanto menor for o valor do AIC menor é a quantidade de informação perdida e, portanto, melhor é o modelo. A medida AIC é definida por (Gomes (2012)):

$$AIC = -2 \cdot [log(L) - k] \tag{3.14}$$

onde k é o número de parâmetros do modelo e L é o valor da verosimilhança para o modelo corrente.

Relativamente aos processos utilizados para seleccionar o melhor modelo, o procedimento mais comum é o método *stepwise* que, de acordo com algum critério (usualmente o AIC) e partindo do modelo saturado (direcção *backward*) ou partindo do modelo nulo (direcção *forward*), escolhe o melhor modelo. Também é possível aplicar o método *stepwise* com a direcção *both* que analisa as duas direcções (*backward* e *forward* em simultâneo.

3.1.2 Coeficientes do modelo

Após o ajustamento do modelo, é usual examinar a contribuição de cada um dos preditores individuais para o ajustamento global. Para tal, é necessário examinar os coeficientes da regressão. Na regressão logística os coeficientes representam a mudança no *logit* por cada unidade de mudança no predictor. No caso de todos os preditores serem variáveis binárias, o coeficiente de cada preditor representa a mudança no *logit* caso o preditor tome valor 1. No entanto, e uma vez que a interpretação do *logit* não é imediata, é usual focar-se no efeito que um preditor tem no *odds ratio*. O *odds ratio* (O.R.) é uma medida que descreve a força da associação (ou da não independência) entre dois conjuntos de observações de variáveis binárias. É usado como uma estatística descritiva e é extremamente importante na

regressão logística, uma vez que, ao contrário do risco relativo, o O.R. trata as duas variáveis a comparar simetricamente. Esta quantidade avalia os *odds* para $Y = 1$ relativamente aos *odds* quando $Y = 0$. Formalmente, sendo p a probabilidade da variável resposta tomar o valor 1 e considerando apenas uma variável regressora X, os O.R. são dados pela equação (3.15) quando a variável regressora é binária e pela equação (3.16) quando a variável regressora é quantitativa.

$$OR = \frac{\frac{\mathcal{P}(Y=1|X=1)}{\mathcal{P}(Y=0|X=1)}}{\frac{\mathcal{P}(Y=1|X=0)}{\mathcal{P}(Y=0|X=0)}} \tag{3.15}$$

$$OR = \frac{\frac{\mathcal{P}(Y=1|X=x+1)}{\mathcal{P}(Y=0|X=x+1)}}{\frac{\mathcal{P}(Y=1|X=x)}{\mathcal{P}(Y=0|X=x)}} \tag{3.16}$$

No caso da variável regressora binária, o *O.R.* representa a mudança no *odds* de $Y = 1$ provocada pela mudança de categoria da variável enquanto que, no caso da variável quantitativa, o *O.R.* representa a mudança no *odds* de $Y = 1$ provocada pelo aumento de uma unidade no valor da variável regressora.

Ao contrário da regressão linear, na qual a importância de um coeficiente é avaliada utilizando um teste t, na regressão logística a significância de um preditor individual pode ser avaliada através do teste da razão de verosimilhanças ou da estatística de Wald.

O teste da razão de verosimilhanças tal como definido na equação (3.13) é também indicado para avaliar a importância que um preditor individual tem para o modelo em estudo (Hosmer (2000), Cohen (2002), Menard (1995)). No caso do modelo com apenas um preditor, basta apenas comparar a *deviance* do modelo corrente com a do modelo nulo segundo uma distribuição de χ^2 com um grau de liberdade. Se o modelo corrente tiver uma *deviance* significativamente menor do que o modelo nulo, pode-se concluir que a associação entre o preditor e a resposta é significativa. Uma vez que há vários *softwares* estatísticos que não têm o teste da razão de verosimilhanças implementado, pode ser mais difícil avaliar a contribuição de um preditor individual, quando a regressão logística é múltipla. Uma solução é ir-se juntando os preditores hierarquicamente, enquanto comparando cada modelo novo com o imediatamente anterior, por forma a determinar a contribuição

de cada novo preditor adicionado ao modelo (Cohen (2002)). Existe, no entanto, muita controvérsia em relação a este procedimento, uma vez que não preserva as propriedades estatísticas nominais e pode ser bastante enganador (Bhandari (2009)).

Alternativamente, quando o objectivo é avaliar a contribuição de um preditor individual para o modelo corrente, pode-se analisar a significância da estatística de Wald. A estatística de Wald, análoga ao teste t na regressão linear, é usada para avaliar a significância dos coeficientes e é dada pela razão entre o quadrado dos coeficientes da regressão e o quadrado do desvio padrão (equivalente à variância) desse mesmo coeficiente. Esta estatística tem distribuição assintoticamente χ^2. Formalmente:

$$W_j = \frac{\beta_j^2}{var(\beta_j)} \tag{3.17}$$

Quando o coeficiente da regressão tem um valor muito elevado, o desvio padrão do coeficiente também tende a ser elevado, o que aumenta a probabilidade de erro de tipo II (Menard (1995)). A estatística de Wald é tendencialmente enviesada quando os dados são escassos, sendo o teste de razão de verosimilhanças mais fiável nestes casos (Agresti (1996)).

3.1.3 Diagnóstico do modelo

Uma vez verificado o ajustamento do modelo e a significância de todos os coeficientes considerados é necessário fazer o diagnóstico do modelo, isto é, é necessário verificar se os pressupostos assumidos inicialmente (como por exemplo a escolha da função de ligação ou da distribuição da variável resposta) são cumpridos, por forma a garantir que as conclusões retiradas estão correctas e a identificar observações mal ajustadas (isto é, que não são bem explicadas pelo modelo). Tal é feito através da análise dos resíduos. Existem várias ferramentas para realizar tal análise, no entanto as ferramentas gráficas são as mais úteis e intuitivas. Formalmente, os resíduos de um modelo correspondem às diferenças entre os valores observados e os valores ajustados:

$$r(i) = y(i) - \hat{y}(i) \tag{3.18}$$

Como a variável resposta é binária (apenas toma os valores 0 ou 1) e como $\hat{y}$ é o *output* do modelo que, tal como já se viu anteriormente, apenas

varia no intervalo $[0,1]$, os resíduos do modelo apenas variarão no intervalo $[-1,1]$, sendo que $r(i) > 0$ corresponde aos casos em que $y(i) = 1$ e que $r(i) < 0$ aos casos em que $y(i) = 0$. Quando $r(i) = 0$ o ajustamento feito pelo modelo é perfeito, isto é, $y(i) = \hat{y}(i)$.

Ao contrário dos modelos de regressão linear, nos quais existe uma componente aleatória à qual se pode associar o valor dos resíduos, no caso dos modelos lineares generalizados, e do modelo de regressão logística em particular, essa componente não existe, fazendo, portanto, mais sentido considerar uma outra definição de resíduos.

Uma alternativa consiste em considerar os chamados resíduos de Pearson que são definidos como (Antunes (2010)):

$$rp(i) = \frac{y(i) - \hat{y}(i)}{\sqrt{\widehat{var}(Y(i))}} \tag{3.19}$$

sendo $\sqrt{\widehat{var}(Y(i))}$ uma estimativa do desvio padrão de $Y(i)$.

Os resíduos de Pearson padronizados são dados por:

$$\frac{rp(i)}{\sqrt{1 - h_{ii}}} \tag{3.20}$$

onde h_{ii} representa o i-ésimo termo da diagonal da matriz de projecção generalizada $\mathbf{H}$ (também chamada *hat matrix* uma vez que $\hat{y} = Hy$), que pode ser calculado através da expressão $h_{ii} = x(i)^T(X^TX)^{-1}x(i)$. A desvantagem dos resíduos de Pearson é que a sua distribuição é, geralmente, bastante assimétrica para modelos não Normais, como é o caso do modelo de regressão logística.

Tal como já referido anteriormente, a análise dos resíduos do modelo é feita, principalmente, com recurso a ferramentas gráficas. Entre os diferentes gráficos disponíveis destacam-se:

- *Scatterplot* dos resíduos - também conhecido como diagrama de dispersão. Permite verificar se os resíduos não apresentam qualquer tipo de padrão, assim como se se encontram bem distribuídos no intervalo [-2,2], sendo que, no mínimo, 95% dos resíduos se devem encontrar neste intervalo;

- Resíduos *versus* valores ajustados - permite avaliar se a variância dos resíduos é não constante (existência de heterocedasticidade). Assume-se que os resíduos são independentes dos valores ajustados, querendo dizer que a correlação entre resíduos e valores preditos deve tomar o valor 0 ;

- *Normal Q-Q plot* dos resíduos - permite avaliar se os resíduos seguem uma distribuição Normal(0,1), por comparação aos quantis teóricos desta distribuição;

- Histograma dos resíduos - permite avaliar a simetria (ou assimetria) dos resíduos, podendo ajudar à detecção de padrões nos resíduos.

A análise dos resíduos é muito importante, no sentido de que permite averiguar a existência de desvios sistemáticos no modelo. No entanto, é também de grande importância averiguar a existência de desvios isolados no modelo, isto é, averiguar a existência de observações que não são bem ajustadas pelo modelo, chamadas observações discordantes. Dependendo da influência que estas observações tenham aquando da estimação dos parâmetros do modelo, a existência de observações discordantes pode ser prejudicial, podendo pôr em causa o desempenho do modelo em análise.

Por forma a detectar se uma observação é, ou não, discordante e/ou influente definem-se os seguintes conceitos:

- *Leverage* - mede a influência que a observação tem nos valores ajustados, sendo um indicativo do grau de influência de uma observação;

- Influência - uma observação é influente se a sua exclusão do modelo produzir alterações significativas nas estimativas dos coeficientes dos parâmetros do modelo.

Observações influentes não têm obrigatoriamente resíduos elevados. De facto, quando uma observação considerada influente apresenta um resíduo elevado, há motivos para desconfiar da bondade do ajuste do modelo uma vez que, sendo o resíduo correspondente à observação elevado, o ajuste desta observação não é o correcto e a observação tem muita influência aquando da estimação dos coeficientes.

Outra forma de encontrar observações influentes (com *leverage* elevada) e mal ajustadas (com resíduos grandes) é através da análise da distância de

Cook. A distância de Cook mede o efeito de retirar uma observação do conjunto de observações usado para fazer o ajustamento do modelo uma vez que observações com resíduos e/ou *leverage* elevados podem distorcer o resultado e a precisão da regressão. Observações com uma distância de Cook elevada são consideradas merecedoras de uma análise aprofundada. A distância de Cook é dada por:

$$D_i = \frac{\sum_{j=1}^{n}(\hat{Y}_j - \hat{Y}_{j(i)})^2}{p\, SQE} \tag{3.21}$$

onde SQE é a soma de quadrados devido ao erro e p é o número de parâmetros ajustados no modelo.

Há diversos critérios para se considerar a distância de Cook elevada. Há autores que defendem que considerar $D_i > 1$ é suficiente (Cook (1982)) enquanto que outros indicam que o valor desta distância deve ser ponderada pelo número de observações usadas para fazer o ajuste do modelo (n), através do critério $D_i > 4/n$ (Bollen (1990)). No âmbito desta dissertação usar-se-á o primeiro critério.

Uma vez terminada a análise dos resíduos é necessário medir a capacidade discriminatória do modelo entre os casos em que a variável resposta toma o valor 0 e os casos em que toma valor 1. Partindo de um conjunto de observações distintas das utilizadas para ajustar o modelo, para as quais se sabe o valor real da variável resposta, faz-se predição utilizando o modelo ajustado. Para cada uma destas observações o modelo dará uma propensão (valor entre 0 e 1) que poderá ser interpretada como a probabilidade de que essa mesma observação apresente valor 1 para a variável resposta.

Como os valores preditos pelo modelo são contínuos, ou seja, estão situados no intervalo $(0, 1)$, é necessário definir um ponto de corte (*cut-off*) para se poder classificar e contabilizar o número de predições positivas (variável resposta toma valor 1) e negativas (variável resposta toma valor 0). Este ponto de corte é tal que, se a propensão dada pelo modelo para uma dada observação for inferior a este valor, assume-se que a observação se encontra na categoria 0 da variável resposta. Contrariamente, no caso em que a propensão dada pelo modelo para uma certa observação seja superior a este valor, considera-se que esta observação se encontra na categoria 1 da variável resposta. Uma vez fixado um *cut-off* é, então, possível dividir as propensões

		Resposta real	
		1	0
Resposta estimada	1	VP	FP
	0	FN	VN

Tabela 3.1: Matriz de confusão para um modelo de regressão logística.

em duas categorias distintas: abaixo do *cut-off* (0) e acima do *cut-off* (1).

Como para este conjunto de observações é conhecido o valor real da variável resposta, é possível compará-lo com a categoria da predição feita pelo modelo. Há, então, quatro cenários possíveis:

- A categoria da predição é 1 e o valor real da variável resposta também é 1. A este grupo dá-se o nome de verdadeiros positivos (VP);
- A categoria da predição é 1 e o valor real da variável resposta é 0. A este grupo dá-se o nome de falsos positivos (FP);
- A categoria da predição é 0 e o valor real da variável resposta também é 0. A este grupo dá-se o nome de verdadeiros negativos (VN);
- A categoria da predição é 0 e o valor real da variável resposta é 1. A este grupo dá-se o nome de falso negativo (FN).

Estes conceitos podem ser esquematizados na tabela 3.1 chamada matriz de confusão.

A partir desta matriz é possível calcular as seguintes medidas:

- Precisão - representa a proporção de predições correctas. Esta medida é altamente susceptível a conjuntos de dados desequilibrados uma vez que não tem em consideração o número de elementos pertencentes a cada categoria. Pode facilmente conduzir a conclusões erradas sobre o desempenho do modelo. É dada por:

$$ACC = \frac{Total\ de\ acertos}{Total\ de\ dados\ no\ conjunto} = \frac{VP + VN}{N} \tag{3.22}$$

  em que N representa o número total de dados no conjunto;

- Sensibilidade - representa a taxa de verdadeiros positivos, isto é, representa a capacidade do modelo de predizer correctamente observações que se encontram na categoria 1 da variável de interesse. É dada por:

$$SENS = \frac{Número\ de\ acertos\ positivos}{Total\ de\ positivos} = \frac{VP}{VP + FN} \tag{3.23}$$

- Especificidade - representa a taxa de verdadeiros negativos, isto é, representa a capacidade do modelo de predizer correctamente observações que se encontram na categoria 0 da variável de interesse. É dada por:

$$ESP = \frac{Número\ de\ acertos\ negativos}{Total\ de\ negativos} = \frac{VN}{VN + FP} \tag{3.24}$$

- Eficiência - é a média aritmética da sensibilidade e da especificidade. Na prática, a sensibilidade e a especificidade variam em direcções opostas na medida de que, geralmente, quando um modelo é muito sensível a positivos (categoria 1 da variável resposta), tende a gerar muitos falsos positivos e vice-versa. Assim, um modelo de decisão perfeito (100% de sensibilidade e 100% de especificidade) raramente é alcançado. É dada por:

$$EFF = \frac{SENS + ESP}{2} \tag{3.25}$$

Uma vez que o *cut-off* pode ser seleccionado arbitrariamente é necessário estudar o efeito da selecção de diversos pontos de corte sobre a saída dos dados. Para cada ponto de corte são calculados os valores da sensibilidade e da especificidade, que poderão ser representados em forma de gráfico, chamado curva ROC, que apresenta no eixo das ordenadas os valores para a sensibilidade e no eixo das abcissas o complementar da especificidade, isto é, o valor $(1 - ESP)$. Um exemplo de uma curva ROC é apresentado na figura 3.1.

Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, porém esta dificilmente será alcançada. Na prática, curvas consideradas boas estarão entre a linha diagonal (classificador aleatório) e a linha perfeita, onde, quanto maior a distância à linha diagonal, melhor o desempenho do modelo. A linha diagonal representa um classificador aleatório, isto é, no qual a probabilidade de ser classificado como positivo é igual à probabilidade de ser classificado como negativo, igual a 0.5.

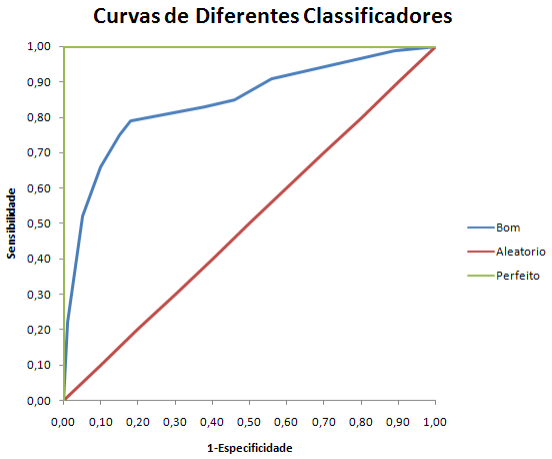

Figura 3.1: Exemplo de uma curva ROC.

Uma medida padrão para a comparação entre modelos é a área sob a curva (AUC - *area under the curve*), que pode ser obtida por métodos de integração numérica, como por exemplo o método do trapézio. Teoricamente quanto maior a AUC melhor o modelo.

Finalmente, a partir de uma curva ROC, deve-se poder seleccionar o ponto de corte que maximiza o desempenho do modelo (*cut-off óptimo*). O *cut-off* óptimo representa um compromisso entre o número de verdadeiros positivos e o número de falsos positivos de forma a que o número de VP seja o maior possível ao mesmo tempo que o número de FP seja o menor possível.

Existem vários métodos para encontrar o *cut-off* óptimo. Nesta dissertação será utilizado o *Youden's index* (usualmente representado por J), que se define como a diferença entre a taxa de verdadeiros positivos e a taxa de falsos positivos. Maximizar este índice permite achar, através da curva ROC, o *cut-off* óptimo. De acordo com a sua definição J é a distância vertical entre a curva ROC e a linha diagonal apresentada na figura 3.1. Se $R(x)$ for a função que descreve a curva ROC, com $x = 1 - ESP$, pode-se escrever $J(x) = R(x) - x$. Quando J é máximo, $J'(x) = 0$, onde J' é a primeira derivada de J, o que se faz com que $R'(x) = 1$, sendo R' a derivada de R. Ora, se $R'(x) = 1$, a tangente à curva ROC é paralela à linha diagonal (que tem declive 1), resumindo-se a tarefa de encontrar o *cut-off* óptimo ao deslize da linha diagonal até que esta se torne uma tangente da curva ROC. Quando

se encontrar a recta tangente à curva ROC obtém-se o par (sensibilidade, 1-especificidade) correspondente ao *cut-off* óptimo.

Capítulo 4

Aplicação

Neste capítulo serão apresentados os vários modelos de regressão logística considerados, assim como a análise do modelo final. Tal como referido na secção 2.3, foram construídos dois modelos distintos: um para a amostra completa e outro para a amostra censurada.

De seguida, e tal como referido anteriormente, foi simulada a utilização do modelo ajustado para a amostra completa. Essa simulação e consequente análise será apresentada na secção 4.3.

4.1 Modelação da amostra completa

Para a construção deste modelo de regressão logística foram utilizadas as U observações disponíveis.

4.1.1 Estratégias de modelação

Uma vez definido teoricamente o modelo de regressão logística, é necessário aplicá-lo ao caso em estudo.

Para tal é necessário dividir os dados disponíveis em dois conjuntos:

- *Training data* - Conjunto de observações usado para ajustar o modelo de regressão logística. Normalmente são usadas 80% das observações. No entanto, uma vez que a dimensão da amostra é bastante elevada, apenas foram usadas 50% das observações.

- *Test data* - Conjunto de observações usado para testar a bondade de ajuste do modelo. Normalmente são usadas 20% das observações. No entanto, e uma vez que só foram usadas 50% das observações para o ajuste do modelo, as restantes 50% foram usadas para testar o modelo ajustado.

Ambos os conjuntos de dados contêm todas as variáveis apresentadas e definidas nas secções 2.2 e 2.3.

Para fazer o ajuste do modelo, usando a *training data*, é utilizada a variável *break* para o mês de Outubro de 2012 como variável resposta, indicando se neste mês existiu quebra de comportamento ou não, e as variáveis *down A*, *down B* e *down C* para os meses Maio, Junho, Julho, Agosto e Setembro de 2012. As variáveis deste modelo, que irá ser referenciado como modelo I, assim como os coeficientes associados e respectivo *p-value*, são apresentadas na tabela 4.1.

Como se pode verificar, há certos coeficientes que não são considerados significativos. Assim, foi utilizado o método *stepwise*, que escolhe o melhor modelo possível, com base no AIC de cada modelo. Neste caso a direcção escolhida foi *backward*, opção que começa com o modelo completo, e que vai retirando variáveis, enquanto vai calculando o AIC do modelo sem essas mesmas variáveis. O modelo resultante desta iteração, apresentada, em parte, na tabela 4.2, será referenciado como modelo II e o seu sumário é apresentado na tabela 4.3.

4.1.2 Diagnóstico do modelo

Um dos pressupostos do modelo de regressão logística é que as variáveis independentes utilizadas para a construção do modelo de regressão sejam não correlacionadas, isto é, a existência de multicolinearidade entre as variáveis pode ser um sintoma de mau ajustamento do modelo. Neste caso, as covariáveis que presidem ao ajustamento do modelo em questão são fortemente correlacionadas, o que já se esperava visto tratarem-se de variáveis que reflectem um comportamento temporal tendencialmente correlacionado (é bastante provável que o valor de uma variável esteja fortemente correlacionado com o valor dessa mesma variável para um mês anterior).

No entanto, como as variáveis explicativas do modelo em questão são variáveis binárias, o método mais correcto para medir o grau de associação

```
> summary(modeloI)

Call:
glm(formula = break_out ~ down_A_set + down_B_set +
    down_C_set + down_A_ago + down_B_ago +
    down_C_ago + down_A_jul + down_B_jul +
    down_C_jul + down_A_jun + down_B_jun +
    down_C_jun + down_A_mai + down_B_mai +
    down_C_mai, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4519  -0.3096  -0.2235   0.3567   2.8033

Coefficients:
                      Estimate   Std. Error  z value   Pr(>|z|)
(Intercept)           -3.67763     0.02548  -144.311    < 2e-16 ***
down_A_set           0.12330     0.03150     3.914   9.06e-05 ***
down_B_set   2.02383     0.02700    74.964    < 2e-16 ***
down_C_set     1.55224     0.03224    48.142    < 2e-16 ***
down_A_ago          -0.00170     0.03294    -0.052   0.958836
down_B_ago   0.50330     0.03133    16.063    < 2e-16 ***
down_C_ago     0.08353     0.03456     2.417   0.015659 *
down_A_jul          -0.14140     0.03359    -4.210   2.55e-05 ***
down_B_jul   0.55525     0.03404    16.309    < 2e-16 ***
down_C_jul    -0.04753     0.03698    -1.285   0.198740
down_A_jun           0.10998     0.03174     3.465   0.000531 ***
down_B_jun   0.66518     0.03501    18.999    < 2e-16 ***
down_C_jun    -0.04747     0.03897    -1.218   0.223196
down_A_mai           0.22511     0.03091     7.283   3.26e-13 ***
down_B_mai   0.79267     0.03470    22.845    < 2e-16 ***
down_C_mai     0.00456     0.03773     0.121   0.903795
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125504  on 99996  degrees of freedom
Residual deviance:  55345  on 99981  degrees of freedom
AIC: 55377

Number of Fisher Scoring iterations: 6
```

Tabela 4.1: *Summary* do modelo I.

```
Step:  AIC=55372.75
break_out ~ down_A_set + down_B_set + down_C_set +
            down_B_ago + down_C_ago + down_A_jul +
            down_B_jul + down_C_jul + down_A_jun +
            down_B_jun + down_A_mai + down_B_mai

                          Df   Deviance   AIC
<none>                          55347    55373
- down_C_jul      1    55350    55374
- down_C_ago      1    55352    55376
- down_A_jun            1    55359    55383
- down_A_set            1    55363    55387
- down_A_jul            1    55366    55390
- down_A_mai            1    55400    55424
- down_B_ago    1    55611    55635
- down_B_jul    1    55615    55639
- down_B_jun    1    55757    55781
- down_B_mai    1    55921    55945
- down_C_set      1    57784    57808
- down_B_set    1    61075    61099
```

Tabela 4.2: *Step* do modelo I.

entre as variáveis é o coeficiente ϕ. O coeficiente ϕ (também conhecido por *"mean square contingency coefficient"*) é uma medida de associação entre duas variáveis binárias e é similar ao coeficiente de correlação de Pearson na sua interpretação. Usando o coeficiente ϕ, duas variáveis binárias são consideradas positivamente associadas se a maioria das observações está distribuída nas células diagonais de uma tabela de contingência 2x2. Contrariamente, duas variáveis binárias são consideradas negativamente associadas se a maioria das observações está distribuída pelas células que não pertencem à diagonal. Neste caso, é apenas possível calcular o coeficiente ϕ para cada par de variáveis constituintes do modelo, o que não tem em conta o efeito das outras variáveis.

Para combater este facto, calculou-se o *variance inflation factor* para as covariáveis do modelo. Na tabela 4.4 é apresentado o valor do VIF para cada uma das variáveis incluídas no modelo II. Como, neste caso, todos os valores do VIF são menores que 5 pode-se afirmar que a multicolinearidade é baixa. Assim sendo, não há indícios da presença de multicolinearidade entre as covariáveis do modelo.

Verificada a parte da multicolinearidade entre as variáveis do modelo é

```
> summary(modeloII)

Call:
glm(formula = break_out ~ down_A_set + down_B_set +
    down_C_set + down_B_ago + down_C_ago +
    down_A_jul + down_B_jul + down_C_jul +
    down_A_jun + down_B_jun + down_A_mai +
    down_B_mai, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4426  -0.3071  -0.2231   0.3567   2.7962

Coefficients:
                      Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)            -3.68071     0.02525  -145.775    < 2e-16 ***
down_A_set           0.12468     0.03117     4.000   6.34e-05 ***
down_B_set   2.02712     0.02674    75.817    < 2e-16 ***
down_C_set     1.54541     0.03147    49.109    < 2e-16 ***
down_B_ago   0.50805     0.03096    16.410    < 2e-16 ***
down_C_ago     0.07632     0.03378     2.259   0.023875 *
down_A_jul          -0.14480     0.03297    -4.392   1.12e-05 ***
down_B_jul   0.55816     0.03380    16.512    < 2e-16 ***
down_C_jul    -0.06364     0.03392    -1.876   0.060632 .
down_A_jun           0.10786     0.03089     3.491   0.000481 ***
down_B_jun   0.64992     0.03177    20.460    < 2e-16 ***
down_A_mai           0.22316     0.03039     7.343   2.10e-13 ***
down_B_mai   0.79134     0.03274    24.170    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 125504  on 99996  degrees of freedom
Residual deviance:  55347  on 99984  degrees of freedom
AIC: 55373

Number of Fisher Scoring iterations: 6
```

Tabela 4.3: *Summary* do modelo II.

Variável	VIF
down A set	1.89711
down B set	1.405509
down C set	1.726248
down B ago	1.860306
down C ago	2.117588
down A jul	2.064078
down B jul	2.238072
down C jul	2.115248
down A jun	1.788032
down B jun	2.003395
down A mai	1.728857
down B mai	2.131451

Tabela 4.4: *Variance Inflation Factors* das covariáveis do modelo II.

necessário fazer a análise dos resíduos do modelo, uma vez que estes são um bom indicador da qualidade do ajustamento.

Um dos pressupostos para o modelo de regressão logística é que os resíduos não apresentem um padrão definido e que 95% dos resíduos estejam no intervalo $[-2, 2]$, uma vez que resíduos elevados (em valor absoluto) são resultado de maus ajustamentos. Neste caso tem-se que cerca de 96% dos resíduos cumprem este requisito. Quanto ao padrão, é visível na figura 4.1 que os resíduos do modelo não apresentam qualquer padrão, não havendo qualquer tipo de heterocedasticidade. É também visível uma grande concentração de resíduos em torno do valor 0, com uma amplitude constante, o que é sinónimo de não haver anomalias no ajustamento do modelo (em termos de função de ligação errada, por exemplo).

A figura 4.2 apresenta a representação gráfica dos resíduos *versus* valores ajustados. As observações (serviços) que apresentam valores mais elevados estão identificadas. Os resíduos destas observações, sendo positivos, correspondem a situações nas quais o serviço apresentou uma quebra no padrão de comportamento, apesar da evidência dada pelo modelo em contrário. No caso do modelo atribuir probabilidade elevada a um serviço de quebrar o seu padrão de comportamento mas na realidade o padrão não ser alterado, os resíduos seriam negativos. No entanto, não há informação na figura 4.2 que evidencie um mau ajustamento por parte do modelo neste caso. A correlação entre os valores ajustados e os resíduos do modelo é 0.205, que é bastante

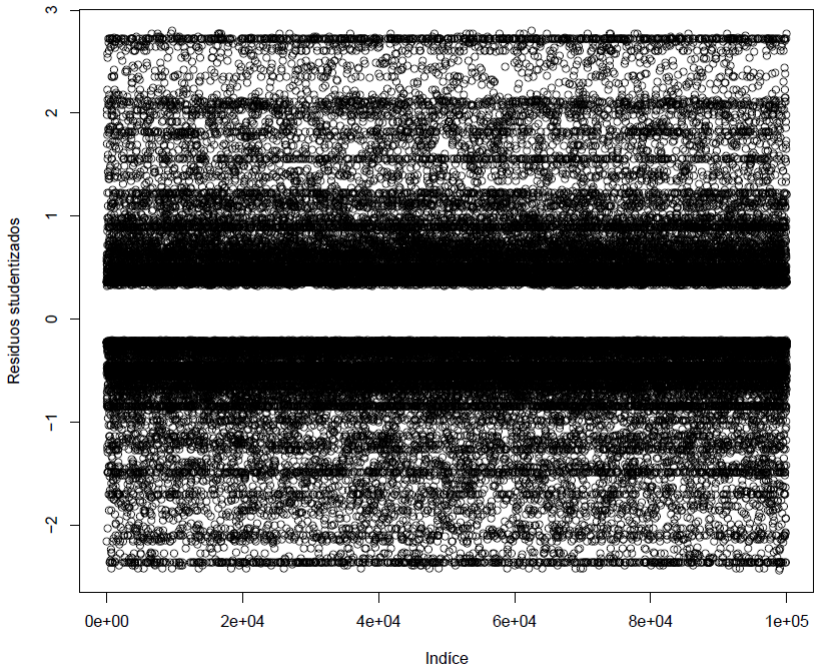

Figura 4.1: Resíduos padronizados do modelo II.

próxima de 0.

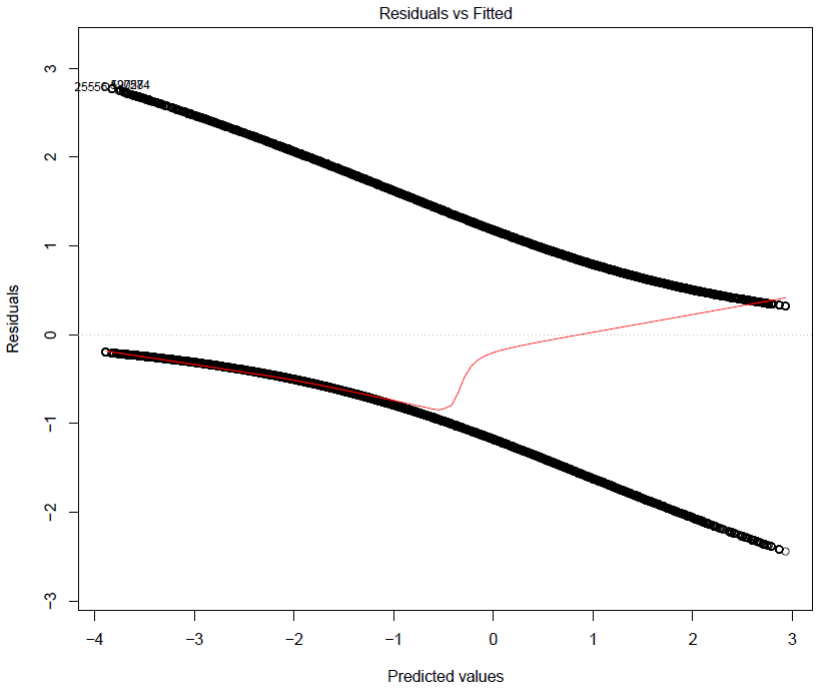

Figura 4.2: Resíduos *versus* valores ajustados do modelo II.

Ao contrário do modelo de regressão linear (simples ou múltipla), o modelo de regressão logística (e mais geralmente o modelo linear generalizado) não exige a normalidade dos resíduos. No entanto, a análise do *Normal Q-Q plot* apresentado na figura 4.3 é útil, no sentido de que permite verificar a estabilidade da variância dos resíduos do modelo. Apesar do ajustamento a uma distribuição Normal não ser perfeito, o *Q-Q plot* da figura 4.3 evidencia uma distribuição subjacente aos resíduos que apresenta, claramente, caudas mais pesadas, quando comparada com uma distribuição Normal. Esta assumpção, assim como a evidência de que a distribuição subjacente aos resíduos seja simétrica, pode ser verificada através histograma da figura 4.4. Ainda na figura 4.3 é bem evidente a existência de dois grupos distintos de observações, devido ao facto da resposta ser binária, o que faz com que os

resíduos tenham dois grandes grupos - o negativo e o positivo, não havendo valores nulos.

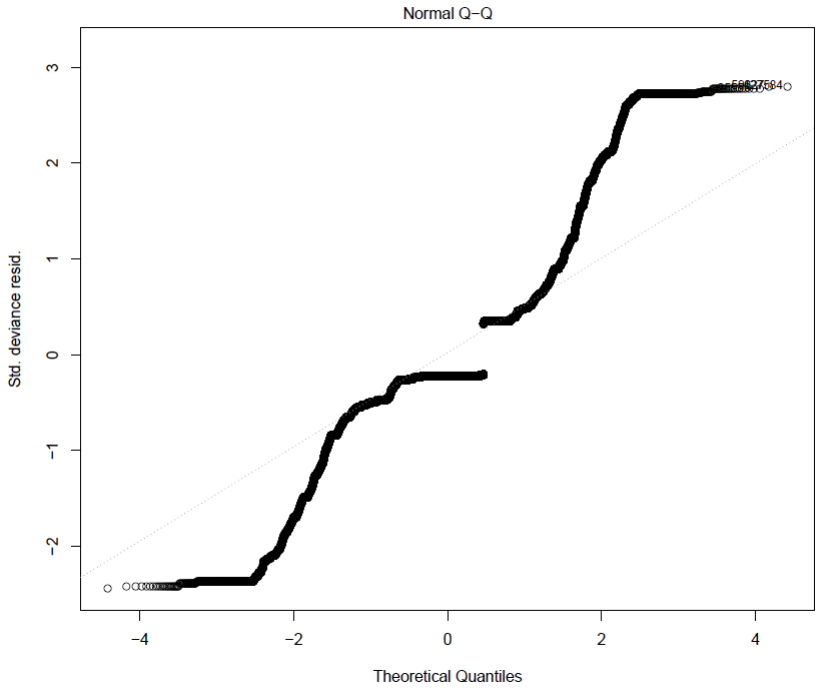

Figura 4.3: Normal Q-Q plot para o modelo II.

O *scale-location plot* do modelo II, apresentado na figura 4.5, mostra a raiz quadrada dos resíduos standardizados (uma espécie de raiz quadrada do erro relativo) como uma função dos valores ajustados e é usado para detectar se a distribuição dos resíduos é constante ao longo do intervalo dos valores ajustados. A análise desta representação gráfica deve conduzir a conclusões semelhantes à análise da representação gráfica dos resíduos *versus* valores ajustados. Tal como na figura 4.2, os valores discordantes estão assinalados.

A figura 4.6 representa o diagrama de dispersão dos resíduos standardizados *versus leverage*. É uma das muitas ferramentas disponíveis para avaliar o ajustamento das observações ao modelo de regressão em uso e é especialmente útil para medir o efeito combinado de observações com *leverage* alta

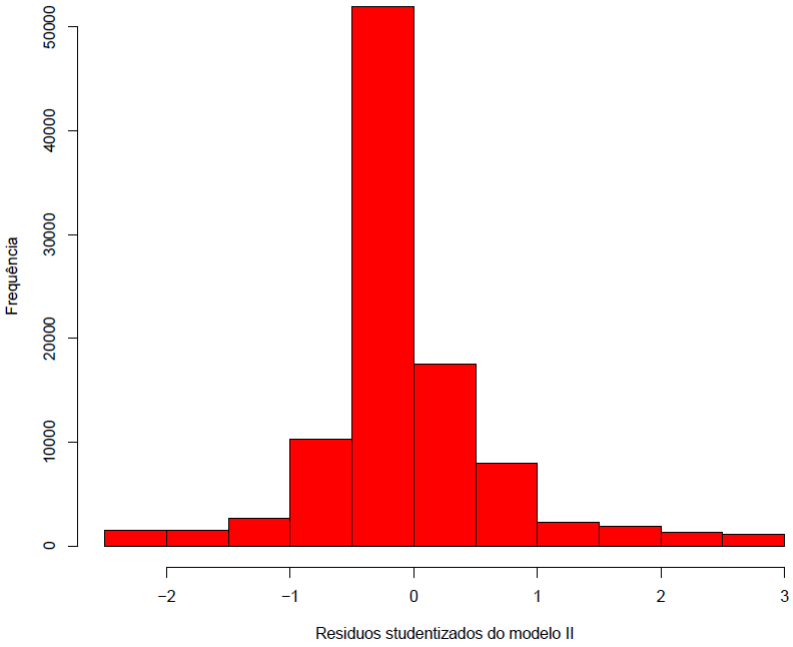

Figura 4.4: Histograma dos resíduos do modelo II.

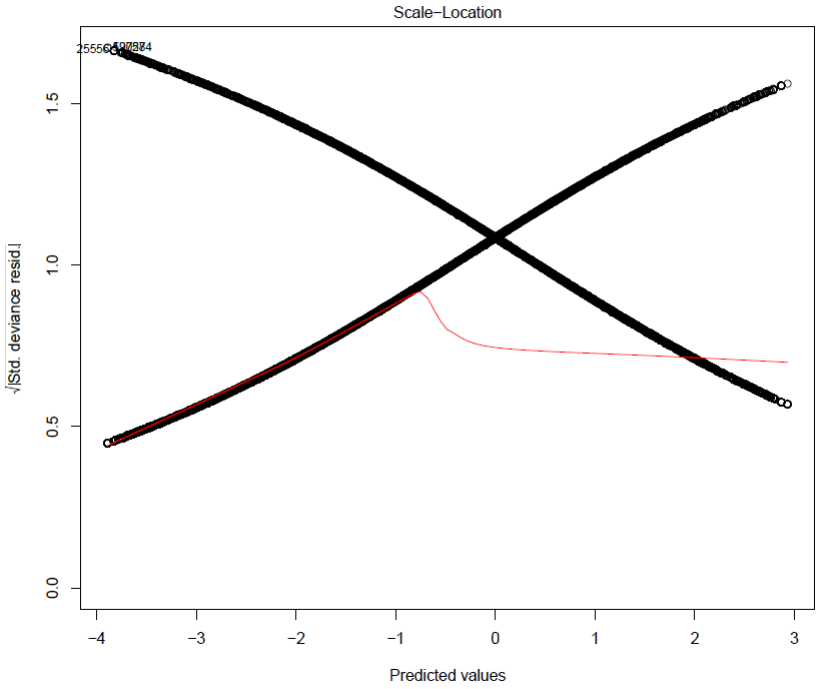

Figura 4.5: Scale-location plot do modelo II.

que possam ser *outliers* da regressão.

Neste caso, todas as observações que apresentam resíduos elevados (e portanto maus ajustamentos, podendo ser considerados *outliers* da regressão) têm *leverage* baixa, significando que não são observações influentes aquando da estimação dos coeficientes da regressão. Por outro lado, todas as observações que apresentam *leverage* alta, sendo, por isso mesmo, influentes, apresentam também resíduos próximos de 0, sendo resultado de bons ajustamentos. Ainda assim, todas as observações apresentam *leverages* muito abaixo de 2 (valor a partir do qual uma observação é considerada influente), não constituindo, portanto, motivo de preocupação.

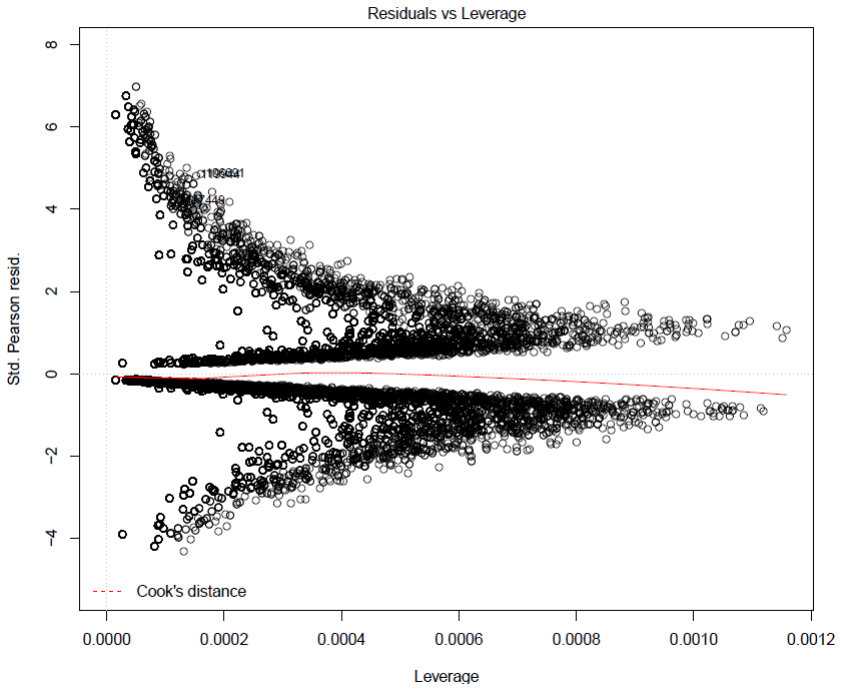

Figura 4.6: Resíduos vs *leverage*.

Outra forma de encontrar observações influentes (com *leverage* elevada) e mal ajustadas (com resíduos grandes) é através da análise da distância de Cook apresentada no gráfico da figura 4.7. Segundo este gráfico, as observa-

ções identificadas são as que apresentam maior distância de Cook, não sendo no entanto suficientemente elevada para se poder ponderar em retirar estas observações da amostra.

A mesma conclusão pode ser retirada do gráfico da figura 4.8, que apresenta a distância de Cook como função da *leverage*. As observações identificadas na figura 4.7 como sendo potencialmente discordantes são também identificadas na figura 4.8. No entanto, e tal como visto anteriormente, todas as observações que apresentam *leverage* elevada apresentam um valor para a distância de Cook baixo (reflectindo também um resíduo próximo de 0) enquanto que as observações que apresentam um valor para a distância de Cook elevado apresentam *leverage* baixa, revelando a baixa influência destas observações aquando da estimação dos coeficientes do modelo.

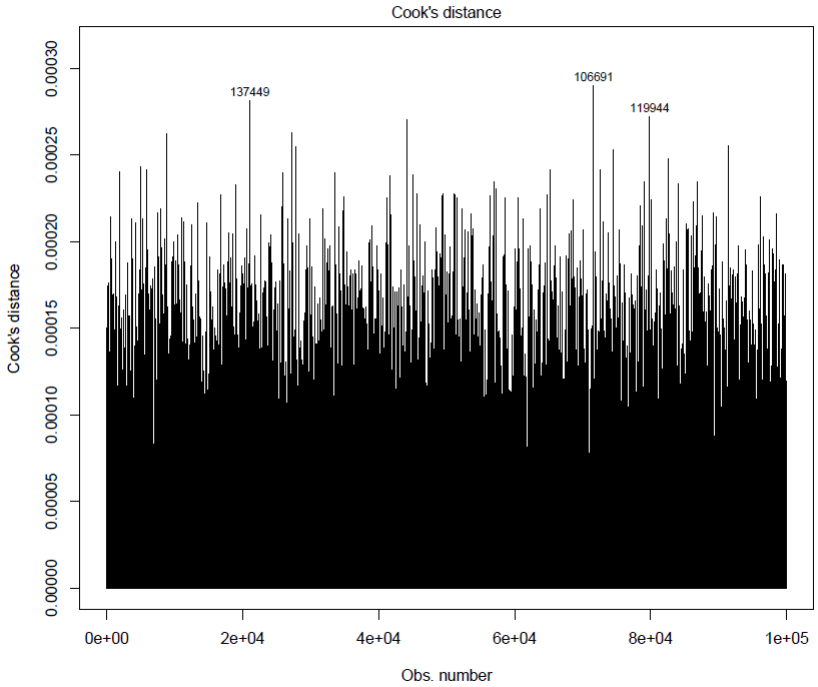

Figura 4.7: Distância de Cook para o modelo II.

Uma vez analisados os resíduos do modelo, é hora de analisar o ajuste do

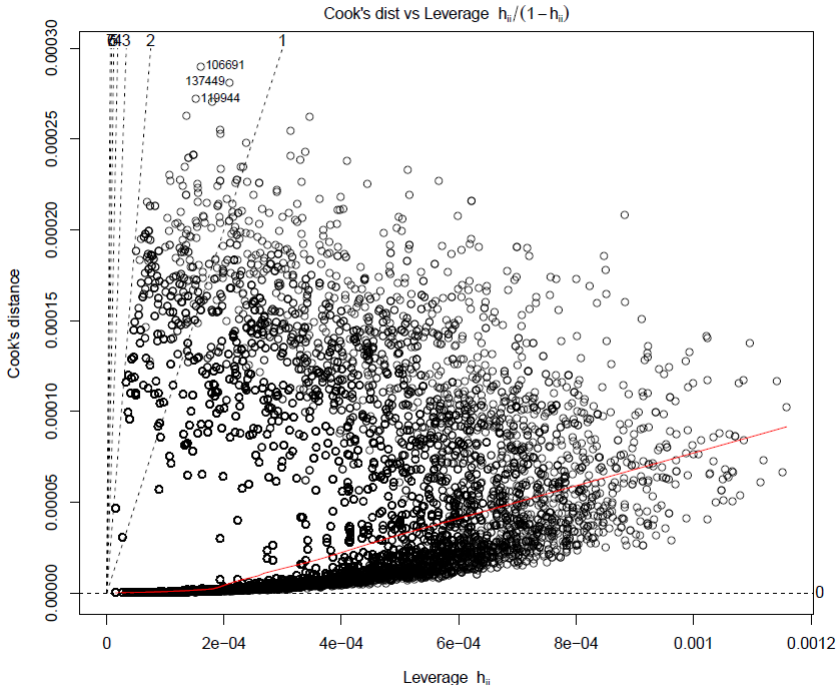

Figura 4.8: Distância de Cook vs *leverage* para o modelo II.

modelo aos dados em questão. Tal é feito, como já referido anteriormente, através da construção e análise da curva ROC e da construção de uma tabela de acertos, uma vez estabelecido o ponto de corte óptimo. A análise do ajuste do modelo é feito com recurso às observações presentes no conjunto *test data*. A curva ROC para o modelo em análise é apresentada na figura 4.9. Uma vez construída a curva ROC é possível encontrar o ponto de corte óptimo. Neste caso usar-se-á este ponto pois quer-se encontrar o melhor compromisso entre a taxa de falsos positivos e a taxa de verdadeiros positivos. No entanto há casos nos quais existe preferência em optimizar uma destas taxas em detrimento da outra.

O método utilizado para encontrar o ponto de corte (*cut-off*) óptimo foi o seguinte:

1. Encontrar, através do método Youden index, o valor óptimo para o par (sensibilidade, 1-especificidade);

2. Uma vez sabendo o valor óptimo para a sensibilidade (0.86) e para a especificidade (0.91), e sabendo que a sensibilidade e a especificidade de um modelo são dadas por, respectivamente:

$$\frac{\#\{\hat{y}=1, y=1\}}{\#\{y=1\}} \tag{4.1}$$

$$\frac{\#\{\hat{y}=0, y=0\}}{\#\{y=0\}} \tag{4.2}$$

   resolver o sistema

$$\begin{cases} \frac{\#\{\hat{y}=1, y=1\}}{\#\{y=1\}} = 0.86 \\ \\ \frac{\#\{\hat{y}=0, y=0\}}{\#\{y=0\}} = 0.91 \end{cases} \tag{4.3}$$

   dado que $\#\{y=0\}$ e que $\#\{y=1\}$ são conhecidos;

3. Uma vez calculadas as quantidades $\#\{\hat{y}=1, y=1\}$ e $\#\{\hat{y}=0, y=0\}$, descobrir, por tentativa e erro, qual o ponto de corte a aplicar ao modelo de modo a que a sensibilidade e a especificidade óptimas do modelo sejam as dadas pela curva ROC.

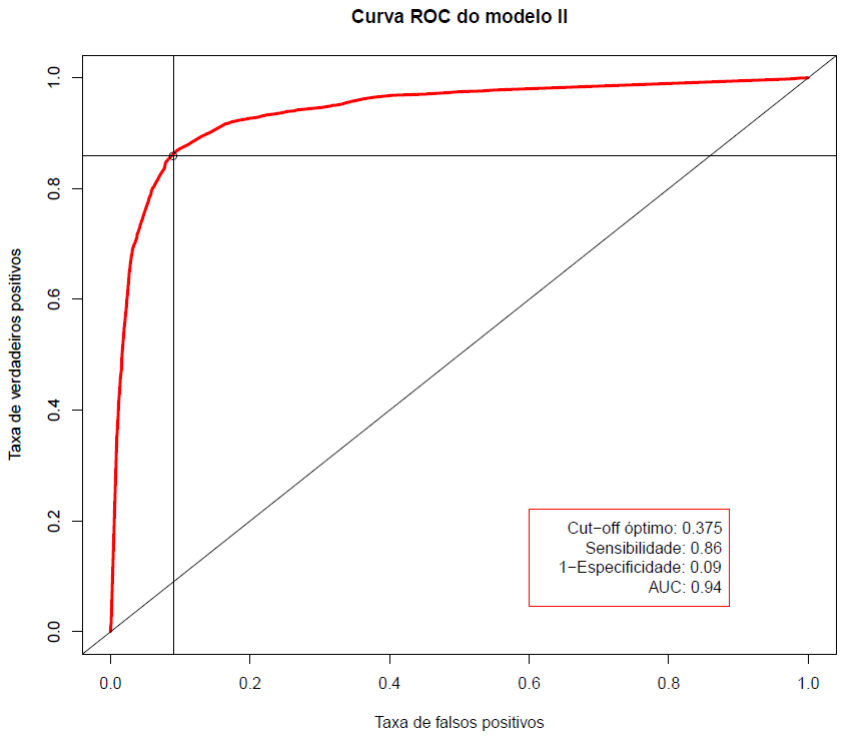

Figura 4.9: Curva ROC do modelo II.

Neste caso, utilizando o ponto de corte 0.375 tem-se que a sensibilidade e que a especificidade do modelo sejam 0.86 e 0.91, respectivamente.

A figura 4.10 mostra, a evolução da taxa de acertos, da sensibilidade, da especificidade e do desempenho do modelo para os distintos pontos de corte possíveis. A vermelho está assinalado o ponto de corte óptimo (0.375) de modo a que possa ser percebido o compromisso que o mesmo representa entre a taxa de verdadeiros positivos e a taxa de falsos positivos.

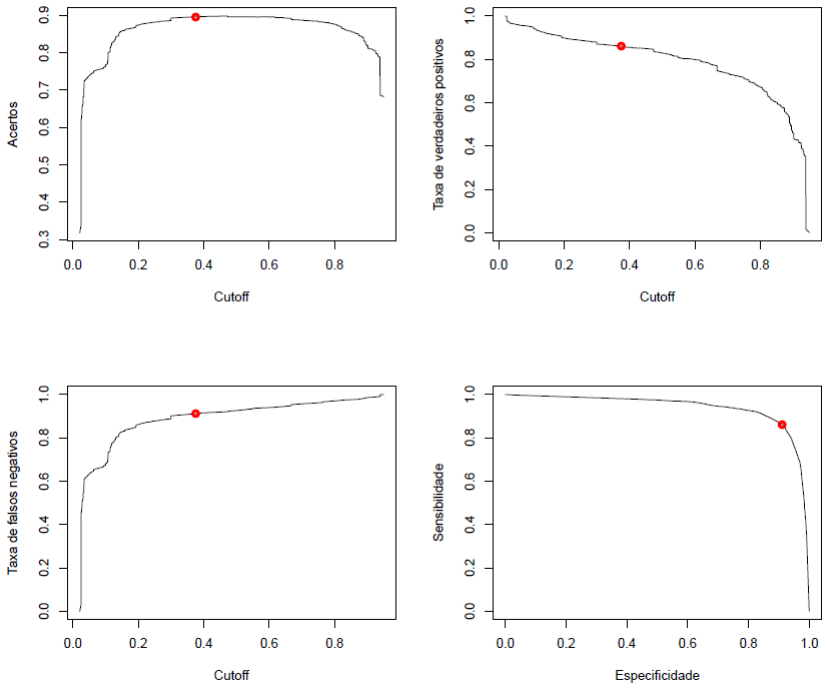

Figura 4.10: Acertos, sensibilidade, especificidade e performance do modelo II.

4.2 Modelação da amostra censurada

Para a construção deste modelo de regressão logística foi utilizada uma amostra censurada. Para a obtenção desta amostra censurada considerou-se

a amostra completa e retiraram-se os serviços que apresentavam o conjunto dos três limites inferiores nulos.

Uma vez que a interpretação da maioria dos resultados é semelhante à apresentada aquando da análise da modelação da amostra completa, a maioria das justificações serão suprimidas.

4.2.1 Estratégias de modelação

Mais uma vez é necessário dividir as observações disponíveis em dois conjuntos de dados: o conjunto *training data* e o conjunto *test data*, cada um formado por 50% das observações que constituem a amostra censurada.

A variável resposta e as covariáveis utilizadas para ajustar o modelo de regressão logística são as utilizadas na secção 4.1.1. As variáveis deste modelo, que irá ser referido como modelo III, assim como os coeficientes associados e respectivo *p-value*, são apresentadas na tabela 4.5.

Como se pode verificar há certos coeficientes que não são considerados significativos, o que faz com que o modelo não seja o melhor possível. Utilizando o mesmo método iterativo que na secção 4.1.1, o modelo resultante do método *stepwise*, método esse parcialmente apresentado na tabela 4.6, será referenciado como modelo IV e o seu sumário é apresentado na tabela 4.7.

É de frisar que, apesar das variáveis explicativas serem as mesmas para o modelo II e para o modelo IV, os coeficientes das mesmas são distintos, não conduzindo obrigatoriamente aos mesmos resultados.

4.2.2 Diagnóstico do modelo

Uma vez que as variáveis utilizadas no modelo IV são as mesmas que as utilizadas no modelo II, não faz sentido analisar novamente a correlação entre as variáveis, uma vez que será igual. Assim, a única medida que realmente depende dos dados utilizados para ajustar o modelo (que são o único critério diferenciador entre o modelo II e o modelo IV) é o *variance inflation factor*, apresentado na tabela 4.8.

```
> summary(modeloIII)

Call:
glm(formula = break_out ~ down_A_set + down_B_set +
    down_C_set + down_A_ago + down_B_ago +
    down_C_ago + down_A_jul + down_B_jul +
    down_C_jul + down_A_jun + down_B_jun +
    down_C_jun + down_A_mai + down_B_mai +
    down_C_mai, family = "binomial", data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4135  -0.4289  -0.2274  -0.2169   2.7937

Coefficients:
                      Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)           -3.64206     0.02577  -141.333    < 2e-16 ***
down_A_set           0.08079     0.03139     2.574   0.010055 *
down_B_set   1.99062     0.02710    73.442    < 2e-16 ***
down_C_set     1.55367     0.03228    48.138    < 2e-16 ***
down_A_ago           0.02442     0.03279     0.745   0.456536
down_B_ago   0.51511     0.03121    16.505    < 2e-16 ***
down_C_ago     0.05727     0.03463     1.654   0.098177 .
down_A_jul          -0.09229     0.03335    -2.767   0.005657 **
down_B_jul   0.55153     0.03391    16.266    < 2e-16 ***
down_C_jul    -0.09625     0.03701    -2.600   0.009309 **
down_A_jun           0.11842     0.03165     3.742   0.000183 ***
down_B_jun   0.67240     0.03479    19.327    < 2e-16 ***
down_C_jun    -0.02132     0.03867    -0.551   0.581514
down_A_mai           0.18534     0.03095     5.987   2.13e-09 ***
down_B_mai   0.74912     0.03450    21.714    < 2e-16 ***
down_C_mai    -0.03003     0.03765    -0.798   0.425144
---
Signif. codes:  0 Ś***Š 0.001 Ś**Š 0.01 Ś*Š 0.05 Ś.Š 0.1 Ś Š 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99402  on 89309  degrees of freedom
Residual deviance: 51058  on 89294  degrees of freedom
AIC: 51090

Number of Fisher Scoring iterations: 6
```

Tabela 4.5: *Summary* do modelo III.

```
Step:  AIC=51086.34
break_out ~ down_A_set + down_B_set + down_C_set +
            down_B_ago + down_C_ago + down_A_jul +
            down_B_jul + down_C_jul + down_A_jun +
            down_B_jun + down_A_mai + down_B_mai

                        Df  Deviance   AIC
<none>                        51060   51086
- down_C_ago     1    51062   51086
- down_A_set       1    51068   51092
- down_A_jul       1    51069   51093
- down_C_jul     1    51071   51095
- down_A_jun       1    51075   51099
- down_A_mai       1    51098   51122
- down_B_jul   1    51331   51355
- down_B_ago   1    51340   51364
- down_B_jun   1    51505   51529
- down_B_mai   1    51568   51592
- down_C_set     1    53494   53518
- down_B_set   1    56632   56656
```

Tabela 4.6: *Step* do modelo III.

Mais uma vez, todos os VIF's são menores do que 5 o que indica baixa multicolinearidade entre as covariáveis do modelo IV.

Em relação aos resíduos do modelo, representados na figura 4.11, 96% dos mesmos encontram-se no intervalo $[-2, 2]$, revelando um bom ajuste. Não é visível qualquer padrão nos resíduos, o que elimina a hipótese de heterocedasticidade, estando, no entanto, muito concentrados em torno do valor 0.

A representação gráfica dos resíduos *versus* valores ajustados é apresentada na figura 4.12. Tal como na figura 4.2, as observações que apresentam resíduos mais elevados estão identificadas e, mais uma vez, apenas correspondem a resíduos positivos. A correlação entre os valores preditos e os resíduos do modelo é 0.178, ainda mais baixa que no modelo II.

Mais uma vez o *Q-Q plot* dos resíduos apresentado na figura 4.13 revela uma distribuição subjacente aos resíduos aparentemente simétrica mas com caudas mais pesadas que uma Normal(0,1). Tal pode ser confirmado no histograma dos resíduos apresentado na figura 4.14.

O *scale-location plot* do modelo IV, apresentado na figura 4.15, mostra

```
> summary(modeloIV)

Call:
glm(formula = break_out ~ down_A_set + down_B_set +
    down_C_set + down_B_ago + down_C_ago +
    down_A_jul + down_B_jul + down_C_jul +
    down_A_jun + down_B_jun + down_A_mai +
    down_B_mai, family = "binomial",
    data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4007  -0.4291  -0.2272  -0.2150   2.7826

Coefficients:
                     Estimate   Std. Error  z value    Pr(>|z|)
(Intercept)           -3.64431     0.02556  -142.565    < 2e-16 ***
down_A_set           0.08610     0.03109     2.769    0.00562 **
down_B_set   1.99521     0.02685    74.321    < 2e-16 ***
down_C_set     1.54617     0.03156    48.997    < 2e-16 ***
down_B_ago   0.52054     0.03085    16.874    < 2e-16 ***
down_C_ago     0.04881     0.03382     1.443    0.14900
down_A_jul          -0.09481     0.03271    -2.899    0.00375 **
down_B_jul   0.55713     0.03365    16.559    < 2e-16 ***
down_C_jul    -0.11114     0.03398    -3.271    0.00107 **
down_A_jun           0.11899     0.03077     3.867    0.00011 ***
down_B_jun   0.67213     0.03158    21.283    < 2e-16 ***
down_A_mai           0.18715     0.03042     6.151   7.69e-10 ***
down_B_mai   0.73596     0.03246    22.674    < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 99402  on 89309  degrees of freedom
Residual deviance: 51060  on 89297  degrees of freedom
AIC: 51086

Number of Fisher Scoring iterations: 6
```

Tabela 4.7: *Summary* do modelo IV.

Variável	VIF
down A set	1.795798
down B set	1.296326
down C set	1.682733
down B ago	1.661392
down C ago	2.047919
down A jul	1.953152
down B jul	2.011153
down C jul	2.052603
down A jun	1.705128
down B jun	1.817897
down A mai	1.670541
down B mai	1.925403

Tabela 4.8: *Variance Inflation Factors* das covariáveis do modelo IV.

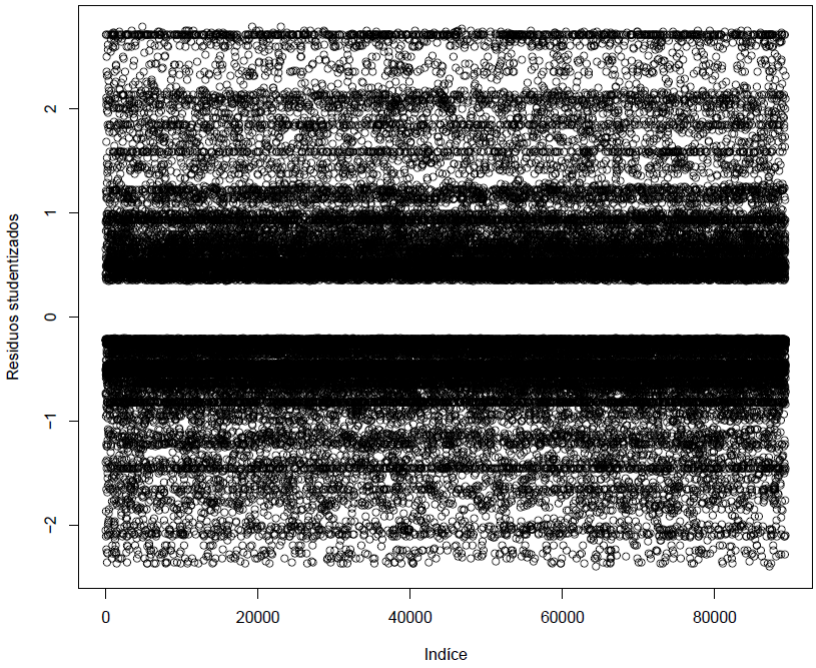

Figura 4.11: Resíduos padronizados do modelo IV.

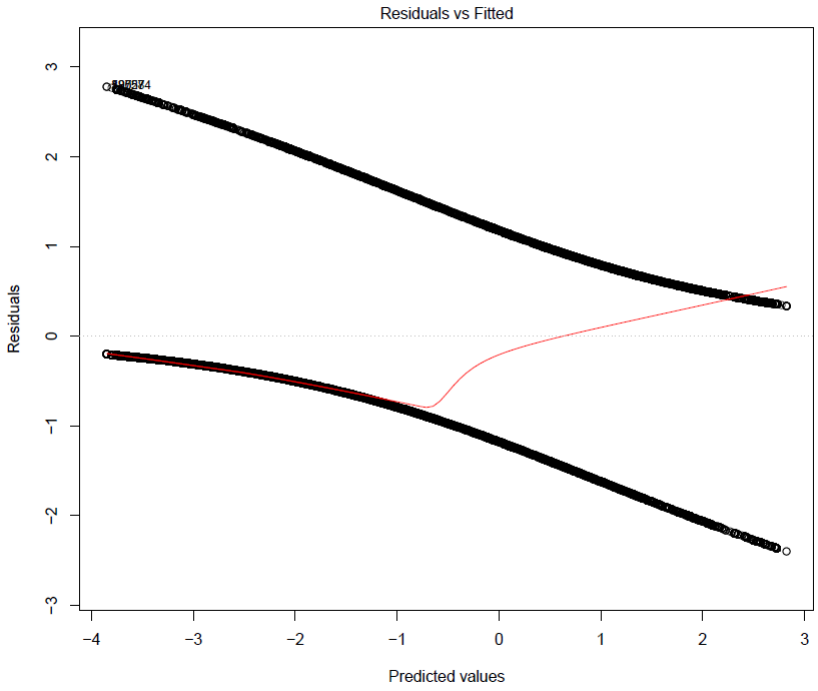

Figura 4.12: Resíduos *versus* valores ajustados do modelo IV.

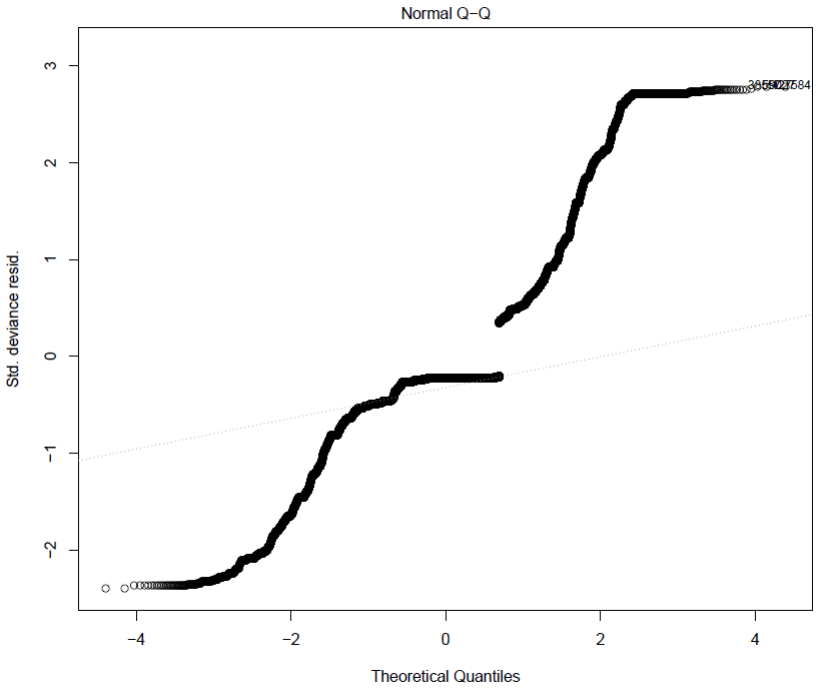

Figura 4.13: Normal Q-Q plot para o modelo IV.

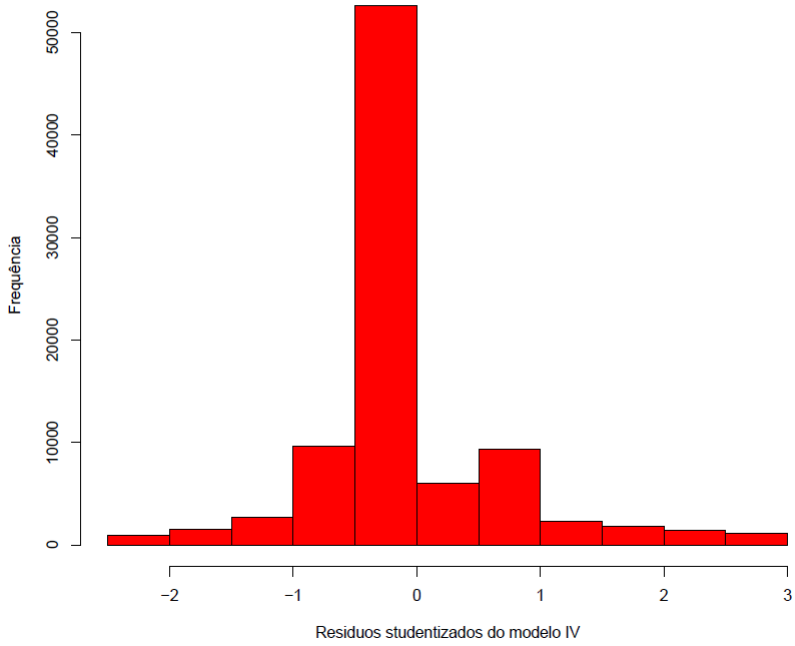

Figura 4.14: Histograma dos resíduos do modelo IV.

que a distribuição dos resíduos é constante ao longo do intervalo dos valores ajustados. A análise desta representação gráfica deve conduzir a conclusões semelhantes que a análise da representação gráfica dos resíduos *versus* valores ajustados. Tal como na figura 4.12, os valores discordantes estão assinalados.

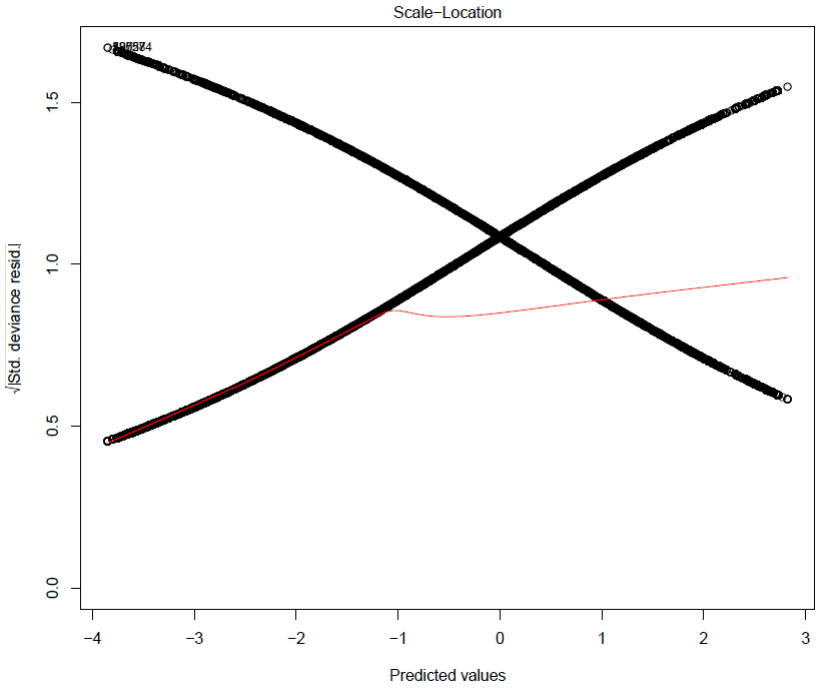

Figura 4.15: Scale-location plot do modelo IV.

Através da figura 4.16 pode-se concluir que todas as observações que apresentam resíduos elevados (e portanto maus ajustamentos, podendo ser considerados *outliers* da regressão) têm *leverage* baixa, significando que não são observações influentes aquando da estimação dos coeficientes da regressão. Por outro lado, todas as observações que apresentam *leverage* alta, sendo, por isso mesmo, influentes, apresentam também resíduos próximos de 0, sendo resultado de bons ajustamentos. Ainda assim, todas as observações apresentam *leverages* muito abaixo de 2 (valor a partir do qual uma observação é considerada influente), não constituindo, portanto, motivos de preocupação.

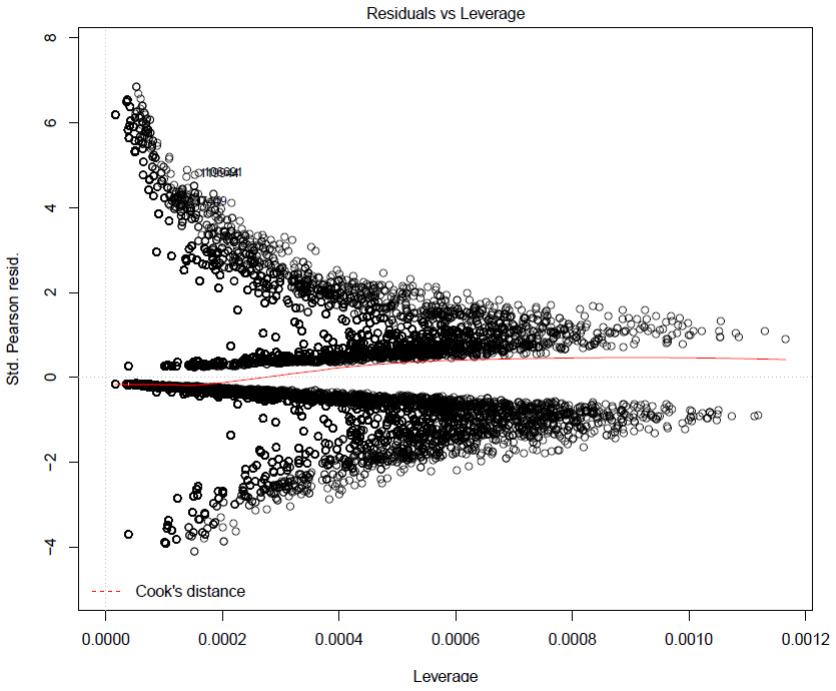

Figura 4.16: Resíduos *versus leverage* para o modelo IV.

Outra forma de encontrar observações influentes (com *leverage* elevada) e mal ajustadas (com resíduos grandes) é através da análise da distância de Cook apresentada no gráfico da figura 4.17. Segundo este gráfico, as observações identificadas são as que apresentam maior distância de Cook, não sendo esta, no entanto, suficientemente elevada para se poder ponderar em retirar estas observações da amostra.

A mesma conclusão pode ser retirada do gráfico da figura 4.18, que apresenta a distância de Cook como função da *leverage*. As observações identificadas na figura 4.17 como sendo potencialmente discordantes são também identificadas na figura 4.18. No entanto, e tal como visto anteriormente, todas as observações que apresentam *leverage* elevada apresentam um valor para a distância de Cook baixo (reflectindo também um resíduo próximo de 0) enquanto que as observações que apresentam um valor para a distância de

Cook elevado apresentam *leverage* baixa, revelando a baixa influência destas observações aquando da estimação dos coeficientes do modelo.

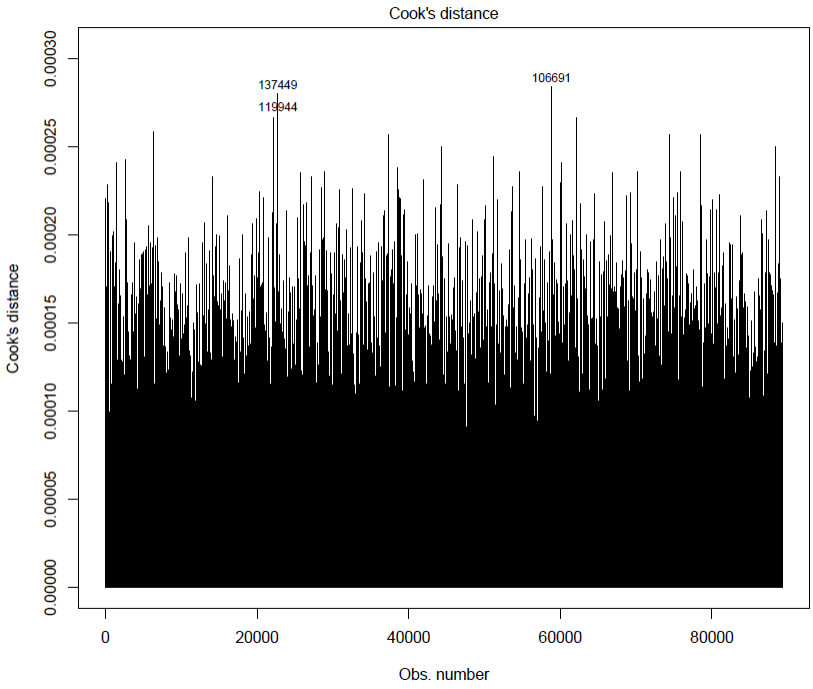

Figura 4.17: Distância de Cook para o modelo IV.

A curva ROC para o modelo IV, apresentada na figura 4.19, permite, mais uma vez, encontrar o valor óptimo para a sensibilidade e para a especificidade do modelo, o que conduz ao cálculo do *cut-off* óptimo.

O método utilizado para encontrar o ponto de corte (*cut-off*) óptimo foi o mesmo utilizado para o modelo II:

1. Encontrar, através do método Youden index, o valor óptimo para o par (sensibilidade, 1-especificidade);

2. Uma vez sabendo o valor óptimo para a sensibilidade (0.84) e para a especificidade (0.88), e sabendo que a sensibilidade e a especificidade

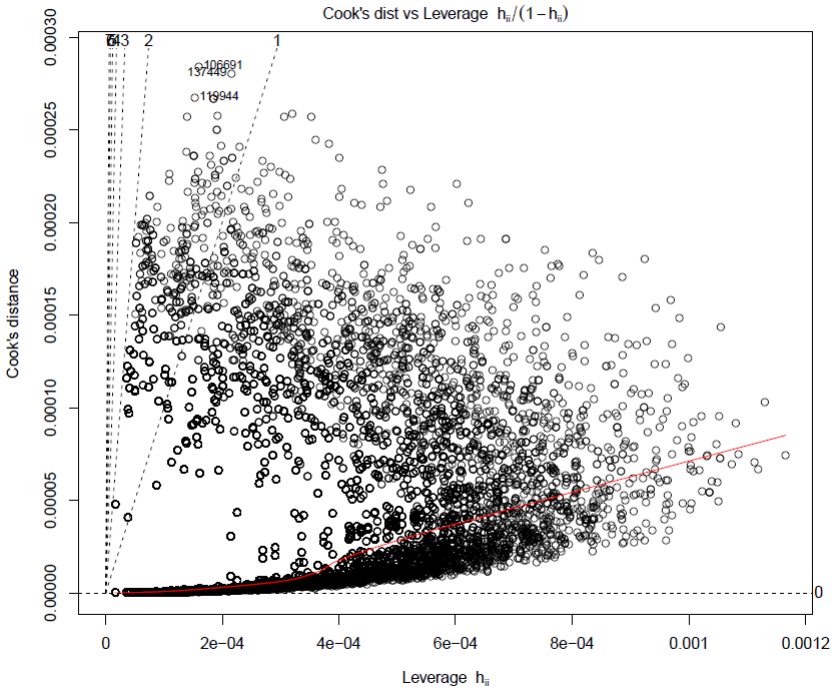

Figura 4.18: Distância de Cook vs *leverage* para o modelo IV.

de um modelo são dadas por, respectivamente:

$$\frac{\#\{\hat{y}=1, y=1\}}{\#\{y=1\}} \tag{4.4}$$

$$\frac{\#\{\hat{y}=0, y=0\}}{\#\{y=0\}} \tag{4.5}$$

resolver o sistema

$$\begin{cases} \frac{\#\{\hat{y}=1, y=1\}}{\#\{y=1\}} = 0.84 \\ \\ \frac{\#\{\hat{y}=0, y=0\}}{\#\{y=0\}} = 0.88 \end{cases} \tag{4.6}$$

dado que $\#\{y=0\}$ e que $\#\{y=1\}$ são conhecidos;

3. Uma vez calculadas as quantidades $\#\{\hat{y}=1, y=1\}$ e $\#\{\hat{y}=0, y=0\}$, descobrir, por tentativa e erro, qual o ponto de corte a aplicar ao modelo de modo a que a sensibilidade e a especificidade óptimas do modelo sejam as dadas pela curva ROC.

Neste caso, utilizando o ponto de corte 0.225 tem-se que a sensibilidade e que a especificidade do modelo sejam 0.84 e 0.88, respectivamente e que a *area under the curve* seja 0.92.

A figura 4.20 mostra, a evolução da taxa de acertos, da sensibilidade, da especificidade e da performance do modelo para os distintos pontos de corte possíveis. A vermelho está assinalado o ponto de corte óptimo (0.225) de modo a que possa ser percebido o compromisso que o mesmo representa entre a taxa de verdadeiros positivos e a taxa de falsos positivos.

4.3 Simulação da utilização do modelo

Apesar de se terem obtido resultados bastante bons aquando da utilização do ponto de corte encontrado para induzir uma partição no conjunto de valores que constituem a predição realizada pelo modelo, pode-se considerar não ser do maior interesse da empresa definir um ponto de corte tão inflexível por duas razões:

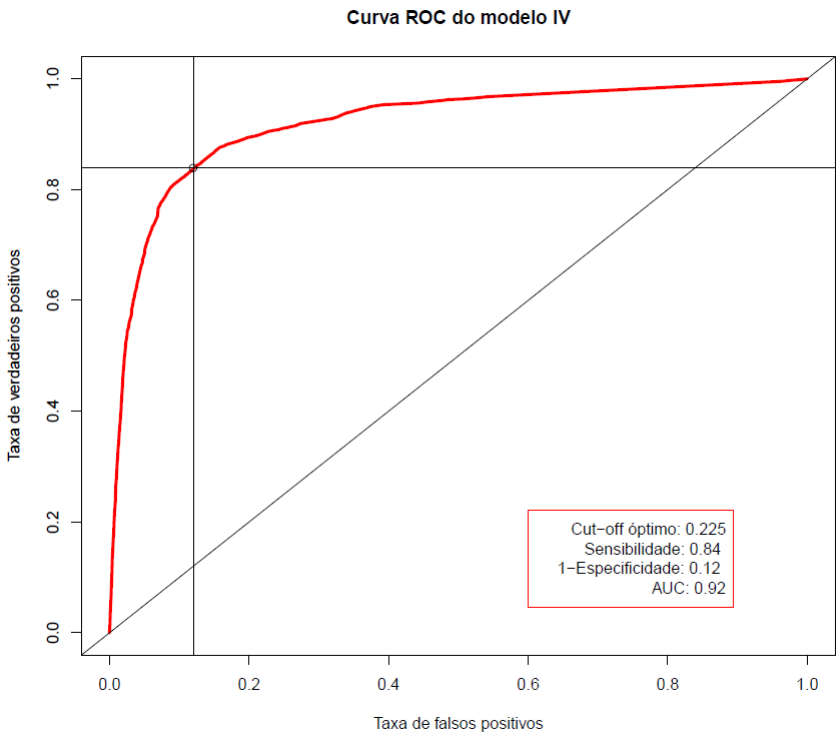

Figura 4.19: Curva ROC do modelo IV.

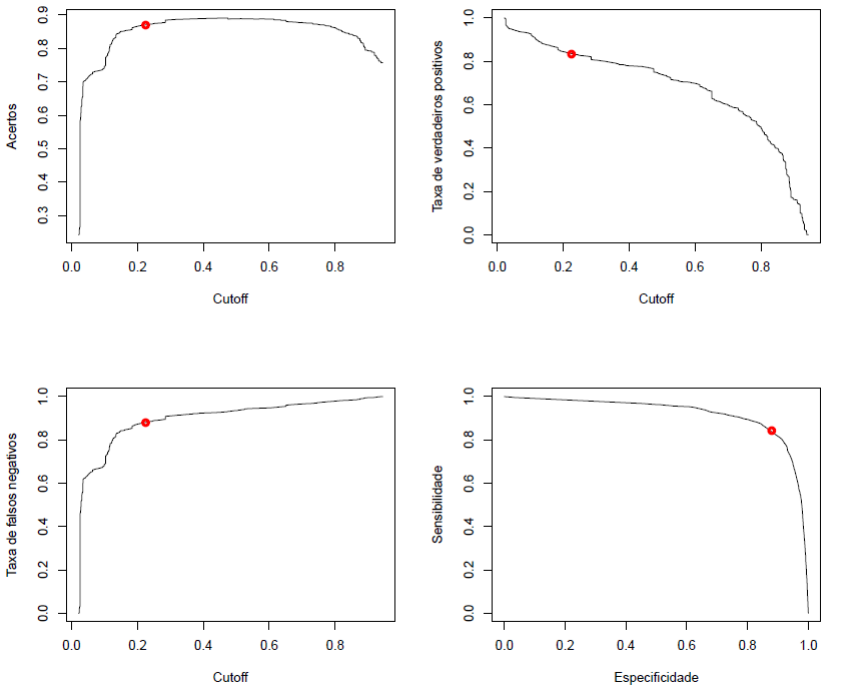

Figura 4.20: Acertos, sensibilidade, especificidade e performance do modelo IV.

Categoria	Propensão	% Serviços	% Não quebras	% Quebras
1	$< 0{,}1$	47 %	97 %	3 %
2	0,1 - 0,2	14 %	88 %	12 %
3	0,2 - 0,3	4 %	76 %	24 %
4	0,3 - 0,4	1 %	64 %	36 %
5	0,4 - 0,5	2 %	48 %	52 %
6	0,5 - 0,6	2 %	49 %	51 %
7	0,6 - 0,7	3 %	36 %	64 %
8	0,7 - 0,8	3 %	32 %	68 %
9	0,8 - 0,9	8 %	13 %	87 %
10	0,9 - 1	16 %	6 %	94 %
	Total	100 %	68 %	32 %

Tabela 4.9: Tabela de predições categorizadas do modelo II.

- O número de serviços presentes na categoria 1 da variável resposta é muito elevado o que impede uma segmentação pormenorizada dos serviços em estudo;
- Dentro dos serviços que o modelo assinala como potencialmente em risco de quebra de comportamento, não há maneira de se identificar quais os clientes que apresentam um risco maior do que os outros.

Assim, no lugar de se encontrar um valor que divida os valores da predição do modelo em dois conjuntos distintos, é possível categorizar estes valores como apresentado na tabela 4.9. Para fazer esta categorização foram utilizadas as predições para o conjunto *test data* fornecidas pelo modelo II.

Como se pode verificar, cerca de 15% dos serviços em análise encontram-se na categoria de risco mais elevada, o que representa um número de serviços muito elevado. Além disso, e tal como já discutido anteriormente, os serviços que se encontram nas categorias de risco mais elevadas podem já estar inactivos há algum tempo (correspondendo, portanto, aos serviços retirados da amostra aquando da construção do modelo IV) o que revelará ineficazes quaisquer campanhas de retenção efectuadas.

Surgem assim duas questões: Quais os serviços que já estão para além do ponto de não retorno e qual será este ponto?

Foi, então, simulada a utilização do modelo II durante sete meses, utilizando dados referentes ao período Janeiro - Dezembro de 2012 e, para cada

mês, foi registada a propensão de quebra dada pelo modelo II para cada serviço em análise, ficando-se assim com uma sucessão de propensões de risco de quebra de comportamento para cada um dos serviços. Categorizando estas propensões tal como anteriormente é possível, utilizando a metologia de cadeias de Markov, calcular a probabilidade de um serviço transitar de uma categoria de risco para outra no espaço de um mês, dois meses, três meses, até n meses. Para o propósito desta análise apenas são analisadas as probabilidades de transição a um, dois e três meses, apresentadas nas tabelas 4.10, 4.11 e 4.12 respectivamente.

Uma cadeia de Markov é um sistema matemático que sofre transições de um estado para outro, entre um número finito (ou contável) de estados possíveis. É um processo aleatório normalmente caracterizado pela ausência de memória:o próximo estado depende apenas do estado corrente e não da sequência de eventos que o precederam. Esta característica é chamada propriedade de Markov.

Formalmente uma cadeia de Markov é uma sequência de variáveis aleatórias $X_1, X_2, X_3, ...$ com a propriedade de Markov que, dado o estado presente, os estados futuros e os estados passados são independentes, tal como definido na equação (4.7).

$$\mathcal{P}[X_{n+1} = x | X_1 = x_1, X_2 = x_2, ..., X_n = x_n] = P[X_{n+1} = x | X_n = x_n] \tag{4.7}$$

Os possíveis valores de X_i formam um conjunto contável S chamado o espaço dos estados da cadeia. As cadeias de Markov são frequentemente descritas através de grafos direccionados, onde os arcos são rotulados com as probabilidades de ir de um estado para os outros estados.

A tabela 4.10 mostra que os serviços que no mês n se encontram na categoria 5 ou superior tendencialmente transitam para categorias de risco igual ou superior no mês $n + 1$ enquanto que os serviços que se encontram nas categorias 1 a 4 tendem a manter-se nas mesmas.

A tabela 4.11 mostra que os serviços que no mês n se encontram na categoria 7 ou superior tendencialmente transitam para as categorias de risco 9 e 10 no mês $n + 2$ enquanto que os serviços que se encontram nas categorias 6 e inferiores tendem a transitar para categorias de risco mais baixas.

		Mês $n+1$									
		1	2	3	4	5	6	7	8	9	10
Mês n	1	0.90	0.07	0.01	0.00	0.01	0.01	0.00	0.00	0.00	0.00
	2	0.24	0.64	0.03	0.01	0.02	0.03	0.02	0.01	0.01	0.00
	3	0.15	0.25	0.23	0.07	0.03	0.05	0.07	0.08	0.06	0.01
	4	0.08	0.20	0.14	0.11	0.08	0.06	0.09	0.08	0.16	0.02
	5	0.13	0.10	0.12	0.06	0.05	0.07	0.25	0.09	0.09	0.04
	6	0.08	0.10	0.12	0.08	0.09	0.03	0.23	0.07	0.19	0.02
	7	0.06	0.05	0.06	0.08	0.02	0.10	0.04	0.35	0.14	0.10
	8	0.02	0.06	0.06	0.03	0.06	0.03	0.09	0.10	0.46	0.09
	9	0.01	0.03	0.03	0.04	0.04	0.03	0.04	0.10	0.28	0.41
	10	0.00	0.00	0.01	0.02	0.00	0.01	0.04	0.02	0.04	0.87

Tabela 4.10: Matriz de probabilidades de transição a um mês.

		Mês $n+2$									
		1	2	3	4	5	6	7	8	9	10
Mês n	1	0.84	0.10	0.01	0.00	0.01	0.01	0.01	0.01	0.01	0.00
	2	0.39	0.43	0.03	0.01	0.02	0.03	0.03	0.02	0.03	0.01
	3	0.26	0.25	0.09	0.04	0.03	0.04	0.06	0.07	0.10	0.05
	4	0.18	0.21	0.08	0.05	0.04	0.04	0.07	0.09	0.14	0.10
	5	0.19	0.15	0.08	0.05	0.03	0.05	0.07	0.13	0.14	0.11
	6	0.15	0.14	0.08	0.05	0.03	0.05	0.07	0.13	0.15	0.13
	7	0.11	0.11	0.07	0.04	0.04	0.03	0.08	0.09	0.25	0.18
	8	0.07	0.10	0.05	0.05	0.04	0.04	0.06	0.10	0.21	0.29
	9	0.04	0.06	0.04	0.04	0.03	0.03	0.06	0.07	0.16	0.49
	10	0.01	0.02	0.02	0.03	0.01	0.02	0.04	0.03	0.06	0.78

Tabela 4.11: Matriz de probabilidades de transição a dois meses.

		Mês $n+3$									
		1	2	3	4	5	6	7	8	9	10
Mês n	1	0.79	0.13	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01
	2	0.47	0.32	0.03	0.02	0.02	0.03	0.03	0.03	0.04	0.03
	3	0.32	0.23	0.05	0.03	0.03	0.03	0.05	0.06	0.10	0.10
	4	0.24	0.20	0.06	0.03	0.03	0.03	0.05	0.07	0.12	0.16
	5	0.24	0.16	0.06	0.04	0.03	0.03	0.06	0.07	0.14	0.18
	6	0.21	0.16	0.06	0.04	0.03	0.03	0.06	0.07	0.15	0.20
	7	0.15	0.12	0.05	0.04	0.03	0.03	0.06	0.08	0.15	0.28
	8	0.11	0.11	0.05	0.04	0.03	0.03	0.06	0.07	0.15	0.36
	9	0.07	0.08	0.04	0.03	0.02	0.03	0.05	0.06	0.12	0.50
	10	0.02	0.03	0.02	0.03	0.01	0.02	0.04	0.04	0.07	0.71

Tabela 4.12: Matriz de probabilidades de transição a três meses.

A tabela 4.12 mostra que os serviços que no mês n se encontram na categoria 7 ou superior tendencialmente transitam para a categoria de risco 10 no mês $n+3$ enquanto que os serviços que se encontram nas categorias 6 e inferiores tendem a transitar para a categoria 1.

Através da análise das matrizes de probabilidades de transição apresentadas pode-se assumir que o ponto de não retorno se situa na categoria 7, querendo dizer que quando um serviço entra na categoria de risco 7 é mais provável que se mantenha nesta categoria (ou que transite para uma categoria de risco superior) do que transite para uma categoria de risco mais baixa, sendo já pouco provável que retome o seu comportamento.

A informação obtida através desta análise será útil na medida de que possibilitará uma segmentação mais eficaz dos serviços a contactar com base na previsão fornecida pelo modelo, quer seja o modelo II ou o modelo IV. A cada mês de utilização do modelo poder-se-á considerar uma abordagem proactiva para os serviços que se encontrem nas categorias 7 ou 8 de risco de quebra de comportamento, por forma a evitar que transitem para uma categoria mais elevada, e uma abordagem retroactiva para os serviços que já se encontrem nas categorias 9 ou 10 de risco de quebra de comportamento.

Capítulo 5

Conclusões e trabalho futuro

Nos capítulos anteriores foi abordado o problema de encontrar uma forma de combinar a informação fornecida pelo comportamento dos consumidores de voz móvel pré-pagos para prever a quebra de actividade num futuro próximo. Ao longo desta dissertação foram apresentadas as técnicas estatísticas utilizadas para resolver este problema assim como a sua aplicação ao caso em estudo.

De facto, o principal objectivo deste trabalho era obter uma forma de fazer uma segmentação eficaz entre os clientes em risco de quebra de actividade e os clientes que não incorriam neste risco, melhorando (ou até complementando) os métodos já existentes, aumentando o aproveitamento e foco das campanhas comerciais já existentes. A implementação deste modelo permitirá à empresa não só identificar os clientes em risco mas também acções melhor direccionadas que as actuais, adequadas a cada tipo de cliente.

Com base no valor das variáveis A, B e C para os últimos cinco meses dos quais se dispõe de informação é possível calcular limites de controlo inferiores, considerando indicativos de quebra de actividade valores de B e de C que sejam inferiores aos respectivos limites inferiores uma vez que a adição do valor de A nesta definição apenas deturpava os resultados.

A quebra (ou não quebra) de actividade foi codificada usando uma variável binária que foi, então, utilizada como variável resposta aquando do ajustamento de um modelo de regressão logística. O modelo foi construído tendo como covariáveis as variáveis *down*. Foi simulada a utilização do modelo II durante sete meses, foram categorizadas as predições dadas pelo modelo em

classes de amplitude 0.1, e a esta sucessão de sete classes para cada serviço foi aplicada a metodologia de cadeias de Markov. Esta análise da sucessão de predições permitiu definir os grupos de risco 7 e 8 como serviços ainda recuperáveis e merecedores de uma abordagem proactiva e os grupos de risco 9 e 10 como merecedores de uma abordagem retroactiva.

Devido à percepção de que os serviços pertencentes aos grupos de risco 9 e 10 já se encontrariam inactivos há algum tempo, foi construído um novo modelo, o modelo IV, apenas com base nos serviços que apresentavam pelo menos um dos limites de controlo inferiores positivo (a esta amostra deu-se o nome de amostra censurada).

Após a realização deste trabalho há diversas conclusões que se podem retirar. Em primeiro lugar foi possível caracterizar a base de clientes de voz móvel pré-pagos em termos de padrões de actividade e inactividade (I ou II) uma vez que se mostrou que os serviços, na sua maioria, alternam entre estados de actividade.

Em segundo lugar ficou evidente que para se caracterizar o comportamento de um serviço de voz móvel pré-pago se podem apenas considerar o valor de A, de B e de C, conferindo às variáveis D, E; F, G, H, I, J, L, M e N um estatuto secundário. Esta afirmação tem como pilares não só os resultados da análise exploratória apresentados no capítulo 2 mas também na análise da bondade de ajuste do modelo II e do modelo IV - se as variáveis D, E; F, G, H, I, J, L, M ou N fossem fundamentais para a caracterização do comportamento tal seria visível aquando do ajuste dos modelos.

Com a simulação da utilização do modelo II (apresentada na secção 4.3) foi possível classificar o grupo de risco 7 como o *ponto de não retorno*, isto é, como o grupo de risco a partir do qual é mais provável transitar para um grupo de risco igual ou superior do que para um grupo de risco inferior. Esta classificação é útil na medida de que é possível quantificar o risco de quebra de padrão de comportamento de cada serviço sem se recorrer a uma separação tão drástica como a resultante da utilização do *cut-off*.

Comparando o modelo II e o modelo IV foi considerado de maior utilidade para a empresa o modelo IV uma vez que se encontra melhor ajustado ao caso da amostra sem os serviços para os quais há pelo menos cinco meses que apresentam um padrão de comportamento residual. Este é o único critério diferenciador entre os modelos apresentados.

Analisando em pormenor os coeficientes das covariáveis do modelo IV, apresentados na tabela 4.7, é de frisar que, se não se considerar o *intercept*, a covariável que apresenta o coeficiente mais alto é a covariável *down B* correspondente ao mês imediatamente anterior ao mês que se pretende prever seguida pela covariável *down C* para o mesmo mês. As restantes covariáveis apresentam coeficientes mais baixos do que os coeficientes destas covariáveis sendo que o coeficiente associado à variável *down* B_j é sempre mais alto do que o coeficiente da variável *down* C_j ou da variável *down* A_j, para cada uma dos meses j considerados. Daqui podem-se retirar duas conclusões:

1. As variáveis *down* correspondentes ao mês imediatamente anterior ao mês que se pretende prever são muito mais influentes que as variáveis *down* para qualquer um dos outros meses considerados, o que de certa forma já era esperado no contexto de serviços pré-pagos;

2. As variáveis *down B* são muito mais influentes que as variáveis *down C* ou *down A*, o que reitera a importância destas variáveis sobre as outras variáveis em estudo.

Como sugestão de trabalho futuro, seria interessante ter oportunidade de verificar a aderência do modelo ao longo do tempo uma vez que é esperado, devido à constante mutação neste ramo de negócio, que os meios pelos quais se realizam as comunicações se alterem (é de esperar que algumas das variáveis consideradas neste estudo como caracterizadoras do comportamento venham a ser substituídas por outras aqui não consideradas).

Estas conclusões mostram que existe um grande potencial na classificação dos clientes de acordo com o seu risco de inactividade. De facto, e após alguns testes, deve-se avaliar a aplicação a toda a base de clientes como método efectivo de reduzir o *churn* entre os clientes pré-pagos de voz móvel de forma significativa.

Bibliografia

Agresti, A. (1996) *An Introduction to Categorical Data Analysis.* John Wiley and Sons, Inc.

Antunes, M. (2010) *Texto de apoio à disciplina de CRM e Prospecção de Dados.* Lisboa: CEAUL.

Bhandari, Mohit. e Joensson, A. (2009) *Clinical Research for Surgeons.* Thieme.

Bollen, Kenneth A. e Jackman, R. W. (1990) Regression diagnostics: An expository treatment of outliers and influential cases. *Modern Methods of Data Analysis* pp. 257–91.

Cohen, Jacob e Cohen, P. (2002) *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences (3rd ed.).* Routledge.

Cook, R. Dennis e Weisberg, S. (1982) *Residuals and influence in regression.* New York: Chapman and Hall.

Gomes, J. (2012) Notas de apoio à disciplina de modelos estatísticos.

Hosmer, David W. e Lemeshow, S. (2000) *Applied Logistic Regression (2nd ed.).* John Wiley and Sons, Inc.

Kutner, Michael H. e Nachtsheim, C. J. (2004) *Applied Linear Regression Models, 4th edition.* England: McGraw-Hill Irwin.

Menard, S. (1995) *Applied Logistic Regression Analysis.* Quantitative Applications in the Social Sciences. Sage Publications.

Myers, R.H. e Montgomery, D. (2002) *Generalized linear models: with applications in engineering and the sciences.* Wiley series in probability and statistics. John Wiley and Sons, Inc.

Richter, Yossi. e Yom-Tov, E. (2010) Predicting costumer churn in mobile networks through analysis of social groups. *SIAM - Society for Industrial and Applied Mathematics* .

Turkman, Maria Antónia e Silva, G. (2000) *Modelos Lineares Generalizados - da teoria à prática.* Lisboa: .